DOCUMENT RESUME

ED 228 114 SO 014 503

AUTHOR	Adams, Robert McC., Ed.; And Others
TITLE	Behavioral and Social Science Research: A National Resource. Part II.
INSTITUTION	National Academy of Sciences - National Research Council, Washington, D.C. Assembly of Behavioral and Social Sciences.
SPONS AGENCY	National Science Foundation, Washington, D.C.
REPORT NO	ISBN-0-309-03297-0
PUB DATE	82
NOTE	611p.; For a related document, see SO 014 617.
AVAILABLE FROM	National Academy Press, 2101 Constitution Avenue, N.W., Washington, DC 20418 ($27.50).
PUB TYPE	Viewpoints (120) -- Books (010) -- Information Analyses (070)
EDRS PRICE	MF03 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	*Behavioral Science Research; Behavior Modification; Cognitive Development; Demography; Economics; Health; Income; Reading Skills; Research Methodology; *Social Science Research; Voting
IDENTIFIERS	Psychophysics

ABSTRACT

Areas of behavioral and social science research that have achieved significant breakthroughs in knowledge or application or that show future promise of achieving such breakthroughs are discussed in 12 papers. For example, the paper on formal demography shows how mathematical or statistical techniques can be used to explain and predict change in population characteristics. A paper dealing with behavior and health discusses several biological processes that link behavior to physical illnesses, including stress and such health-impairing habits as smoking, heavy drinking, poor diet, and lack of exercise. In another paper a social anthropologist explores the relationships between culture, social culture, personality, and experience. The remaining papers treat voting behavior research, the life-span approach, income inequality, advances in methods for large-scale surveys and experiments, psychophysics, reading as a cognitive process, property and territoriality, cognitive development in the first years of life, and behavior therapy. (RM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED228114

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.
☒ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
J. Olson
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Part II

Behavioral and Social Science Research: A National Resource

Robert McC. Adams, Neil J. Smelser,
and Donald J. Treiman, editors

Committee on Basic Research in the
Behavioral and Social Sciences

Commission on Behavioral and Social Sciences
and Education

National Research Council

SO 014 503

NATIONAL ACADEMY PRESS
Washington, D.C. 1982

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This report has been reviewed by a group other than the authors according to procedures approved by a Report Review Committee consisting of members of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

The National Research Council was established by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and of advising the federal government. The Council operates in accordance with general policies determined by the Academy under the authority of its congressional charter of 1863, which establishes the Academy as a private, nonprofit, self-governing membership corporation. The Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in the conduct of their services to the government, the public, and the scientific and engineering communities. It is administered jointly by both Academies and the Institute of Medicine. The National Academy of Engineering and the Institute of Medicine were established in 1964 and 1970, respectively, under the charter of the National Academy of Sciences.

Library of Congress Catalog Card Number 82-81776

International Standard Book Number 0-309-03297-0

Available from

NATIONAL ACADEMY PRESS
2101 Constitution Avenue, N.W.
Washington, D.C. 20418

Printed in the United States of America

COMMITTEE ON BASIC RESEARCH IN THE BEHAVIORAL AND SOCIAL SCIENCES

ROBERT McCORMICK ADAMS (*Chair*), The Oriental Institute, University of Chicago

*WILLIAM J. BENNETT, National Humanities Center, Research Triangle Park, North Carolina

DAVID A. HAMBURG, John F. Kennedy School of Government, Harvard University

JUANITA M. KREPS, Durham, North Carolina

GARDNER LINDZEY, Center for Advanced Study in the Behavioral Sciences, Palo Alto, California

ELIZABETH F. LOFTUS, Department of Psychology, University of Washington, Seattle

JAMES G. MARCH, Graduate School of Business, Stanford University, and Hoover Institution, Stanford, California

JESSICA T. MATHEWS, Editorial Board, *The Washington Post*

PHILIP MORRISON, Department of Physics, Massachusetts Institute of Technology

CHARLES A. MOSHER, Washington, D.C.

KENNETH PREWITT, Social Science Research Council, New York

†PETER H. ROSSI, Department of Sociology, University of Massachusetts

PAUL A. SAMUELSON, Department of Economics, Massachusetts Institute of Technology

NEIL J. SMELSER, Department of Sociology, University of California, Berkeley

SAM BASS WARNER, JR., Department of History, Boston University

*Resigned December 1981

†Resigned July 1981

ROBERT J. ZAJONC, Department of Psychology, University of Michigan

DONALD J. TREIMAN, Study Director
PATRICIA A. ROOS, Research Associate
ROSE S. KAUFMAN, Administrative Secretary

Contents

INTRODUCTION 1

RESEARCH IN FORMAL DEMOGRAPHY
Jane Menken and James Trussell 6

THE STUDY OF VOTING
Philip E. Converse, Heinz Eulau, and Warren E. Miller 33

BEHAVIOR AND HEALTH: THE BIOBEHAVIORAL PARADIGM
David S. Krantz, David C. Glass, Richard Contrada, and Neal E. Miller 76

EARNINGS AND THE DISTRIBUTION OF INCOME: INSIGHTS FROM ECONOMIC RESEARCH
James J. Heckman and Robert T. Michael 146

CULTURAL MEANING SYSTEMS
Roy G. D'Andrade 197

THE LIFE-SPAN PERSPECTIVE IN SOCIAL SCIENCE RESEARCH
David L. Featherman 237

ADVANCES IN METHODS FOR LARGE-SCALE SURVEYS AND EXPERIMENTS
Judith M. Tanur 294

RESEARCH IN PSYCHOPHYSICS
L. D. Braida, Tom N. Cornsweet, N. I. Durlach, David M. Green, Herschel Leibowitz, Alvin Liberman, R. Duncan Luce, Richard Pew, and Carl Sherrick 373

READING AS A COGNITIVE PROCESS
Patricia A. Carpenter and Marcel Adam Just 406

TERRITORY, PROPERTY, AND TENURE
Robert McC. Netting 446

COGNITIVE DEVELOPMENT IN THE FIRST YEARS OF LIFE
Katherine Nelson 503

FROM EXPERIMENTAL RESEARCH TO CLINICAL PRACTICE: BEHAVIOR THERAPY AS A CASE STUDY
G. Terence Wilson 554

CONTRIBUTORS 603

Introduction

The Committee on Basic Research in the Behavioral and Social Sciences was established in 1980 by the Assembly of Behavioral and Social Sciences, now the Commission on Behavioral and Social Sciences and Education, at the request of the National Science Foundation. As its first task, the committee was asked to assess the value, significance, and social utility of basic research in the behavioral and social sciences.

Part I of the committee's report, published as a separate volume, is an assessment of the variety of ways in which basic research in the behavioral and social sciences contributes to society. Following an initial review of the subject matter and modes of research in the behavioral and social sciences, the report illustrates the nature of research in these fields through a series of brief vignettes describing significant scientific advances. It goes on to describe some illustrative applications of knowledge gained from behavioral and social science research, ranging from the diffuse impact of new knowledge on the way we think about ourselves and our society, to improvements in the procedures for collecting and analyzing information for planning and policy formation in both the public and private sectors, to direct applications for the development and improvement of goods and services that contribute to individual and collective well-being. The report concludes that basic research in the behavioral and social sciences is a valuable national resource that warrants full public support.

Part II is a supplement to the committee report. It consists of a set of papers that illustrate in detail areas of behavioral and social science research that have achieved significant breakthroughs in knowledge or application or that show promise of achieving such breakthroughs in the near future. The committee commissioned these authors to try to convey the depth, richness, and excitement of the intellectual activity in selected areas of ongoing research. We think they have succeeded.

The range and diversity of research in the behavioral and social sciences make a comprehensive survey impossible. We have had to select a mere handful of topics for examination; the selection process was aided by consultation with many behavioral and social scientists but subject to the usual practical contingencies of author availability. The choice of topics, and the disciplines and fields they represent, should not be regarded as a sampling of all the work in the behavioral and social sciences, but the choice does exemplify their range and quality.

In their paper on formal demography Menken and Trussell show how demographic techniques, which are largely mathematical or statistical in order to achieve precision, have been directed toward measuring, explaining and predicting change in population characteristics. Demographic models provide insights into population problems involving fertility, mortality, age structure, migration, and population distribution in countries around the world.

Research on voting behavior, which began 50 years ago with the collection of survey data on how individual voters made their decisions, has been a development of successively more complex--and more realistic--models of voting behavior. Converse et al. explain how the "simple act of voting" is not simple at all, recounting some of the surprises that researchers in this field have found.

As the effects of behavior on biological processes have come to be recognized, the behavioral and social sciences have made increasingly significant contributions to understanding problems of physical health. The integration of behavioral and biomedical knowledge is a research area of extremely rapid development. Krantz et al. discuss several processes that link behavior to physical illness of various kinds, including stress in particular and such health-impairing habits as smoking, heavy drinking, poor diet, and lack of exercise.

Few topics in the social sciences have received more attention than the nature and causes of the distribution of income. Heckman and Michael describe the extent of income inequality in the United States and review the extensive research literature by economists that attempts to explain variations in the earnings of individuals, the major source of income.

Culture is the concern of social anthropologists, and D'Andrade explores the relationships between culture, social structure, personality, and experience. Meaning systems, on which a widely used concept of culture is based, embody the categories with which we perceive the world, that influence our conscious or unconscious choices among alternative courses of action, and that help to evoke many of our feelings. Marriage and success are two cultural entities he uses to illustrate this cognitive approach to culture.

A new insight that informs much recent research in branches of psychology, sociology, and history is the recognition that developmental changes in behavior and personality occur throughout life, among adults and the elderly as well as among infants and children. This life-span perspective is useful to many different kinds of investigations of human development. Featherman uses current research on social stratification and mobility, psychometric intelligence, and the history of the family to illustrate the ways in which the life-span approach can illuminate stability and change in behavior and personality.

Statistical techniques of collecting, analyzing, and interpreting data are the building blocks of the behavioral and social sciences, and the technology of survey research is one of the most valuable. Tanur discusses some of the methods used in large-scale surveys and social experiments; how they have been refined; and the uses of these methods in government, commercial polling firms, market research, and university-based research institutes. The research she describes suggests that far from being a constant set of known tools, survey methodology is an active field undergoing constant innovation and revision.

How individuals perceive stimuli that impinge on the senses is the focus of psychophysics, a discipline that studies the relationships between the material and the mental worlds. Braida et al. describe current psychophysical research on human visual and auditory systems. They also give an account of some of its attempted applications, including the reduction of nighttime driving accidents, the development of reading machines for the blind, and the correction of misleading visual information in landing approaches of airplanes at night. Success in terms of solving practical problems in these areas is only partial, but the promise as well as the need is evident.

Our understanding of cognitive development in infants and young children has increased greatly as a result of basic research. Nelson traces this line of research, noting developments that have important implications for social policy, such as whether day care has negative effects on children's intelligence. Researchers have learned much about the interaction of biological and environmental forces and are still discovering the details of what conditions are necessary for normal development.

One interest of cultural anthropologists is the study of behavior and beliefs concerning the ownership of land in the context of larger economic and social systems. While politics and law often isolate land tenure from other cultural factors, recent research in anthropology has insisted that ecological relationships--including the physical environment, agricultural uses, labor expenditure, technology, market forces, and historic ideologies--must all be considered. Netting illustrates these themes with ethnographic case studies of hunter-gatherer territoriality, fishing rights, and landholding among shifting and intensive agriculturalists. Insights from this body of work can help us anticipate the probable success of land reform efforts in developing nations.

Basic research in several disciplines has led to a new interest in reading as a cognitive process. Recently there has been considerable success in formalizing what is known about the structure of language, the nature of the perceptual process in reading, and ways to facilitate the learning of this essential skill. Carpenter and Just discuss current basic research on reading as a thinking process that can be taught and improved.

Based on discoveries from experimental research on learning, behavior therapy is a new approach to the assessment and treatment of emotional and behavioral disorders. Wilson's paper gives examples of the wide range of problems currently addressed by behavioral treatment methods. These methods include fear reduction techniques for the treatment of anxiety disorders, positive reinforcement of desired behavior in psychotic patients, and training in self-control to induce behavior change. In this area there has been a systematic progression from experimental research to general clinical application.

These papers illustrate a number of themes that characterize basic research in the behavioral and social sciences in general. Advances are continual, making progress in many directions possible. As questions are asked, methods are found to answer them; as new methods are applied, new questions emerge. There is a steady interaction of theory and method, the substantive and the technical, that works to expand the base of knowledge on which the behavioral and social science disciplines rest. And results beget new questions, too; the process is incremental. Successive refinements of models and theories add complexity and depth to our understanding of social structures and processes, for instance, how labor markets operate or voters vote. This volume is a portfolio of accounts of current research in a number of behavioral and social science fields; the overall activity is constant, growing, and exciting.

Research in Formal Demography

Jane Menken and James Trussell

WHAT IS FORMAL DEMOGRAPHY?

Demography is a discipline that combines some of the quantitative nature of the physical sciences with the direct attention to human problems that characterizes much of the social sciences. Its connection to obviously countable or measurable entities--such as age or duration of marriage or the number of persons, births, and deaths--permits the precise use of mathematics, often rather complex. At the same time, the importance of human mortality and fertility and of demographic stability and change to human welfare leads demographers to a strong natural interest in the application of results of basic research to real-world problems.

As in the natural sciences, basic research in demography, even the most abstract, ultimately yields information of great practical use, while attempts to solve practical problems uncover new puzzles or new information that is useful for further basic research. Application of the understanding of population structure gained from the most abstract, most mathematical demographic research has become pervasive, to the extent that both the 1980 and 1981 budgets of the United States government contain 5 to 10 pages devoted to the long-term outlook for age composition and fertility in the United States and their implications for future economic and social problems.

Some results of demography are cautionary in nature, in that they run counter to intuition. A simple proposition may serve as an example: A healthy country always has a lower death rate than a less healthy one--true or false? The United States in 1977 had a mortality rate of 9 per 1,000 population, about 80 percent higher than that of Taiwan (at 5 per 1,000). Yet few would accept the proposition that people in the United States are less healthy than people in Taiwan. What explains the discrepancy? Demographic research has shown that it is crucial to consider the age structure and mortality by age. At every age, people in the United States have lower death rates than those in Taiwan. However, more Americans are older and are therefore in higher-risk age groups than are the Taiwanese. The Taiwanese age structure is so much more concentrated in the young age groups, which in every country have relatively low mortality (except at the very youngest ages), that its overall death rate is considerably lower than the U.S. rate. This kind of understanding of demographic processes has led to the development of measures of mortality that are free of the contamination of the effects of age structure.

Demographic research is directed toward measuring, explaining, predicting, and in some cases influencing change in population characteristics--primarily fertility, mortality, age structure, and migration or population distribution. As in most of the natural sciences, descriptive and methodological studies have been and are crucial. How does fertility vary around the world? Are there identifiable patterns by age? By region? How does one measure fertility (or mortality) in areas for which no records of births (or deaths) are kept? To what characteristics of the population or individual or couple is fertility related? With what characteristics is fertility change associated? Are there useful predictors of future demographic change? Although many, perhaps most, of these questions cannot yet be answered completely, basic research through demographic model building or formal demography has played a significant role in producing insights and techniques for more applied work as well as for population policy.

DEMOGRAPHIC MODELS

Demographic models have been primarily of two types--those that are causal and those that are purely descriptive. Causal models are useful in examining the interaction between one or more factors (for example, birth rates and death rates) or their effect on a variable that is taken as dependent. Descriptive models usually attempt to summarize typical patterns, often age patterns, of occurrence of some demographic event. These models are based on observed empirical regularities in the occurrence of deaths, births, and marriages. They attempt not to examine what causes these age patterns but rather to describe them.

It is reasonable to ask whether there are useful applications of such descriptive models, beyond the interest of basic research in discovering that there are commonalities, or empirical regularities, in demographic processes in many different populations. The models have been applied both in demographic research and in work that has implications beyond the world of demographic research. If a model offers a concise description of some aspect of fertility, mortality, or other demographic process, the characteristics of any population can be summarized by the parameters of the model. This shorthand description facilitates the study of variation among populations or within a population over time. Demonstrating that certain populations share common characteristics may also provide a clue toward elucidating the factors that cause them to be similar. In addition, the descriptive models have been used to adjust data from populations to which observations are subject to error. By using the observations to estimate the parameters of the model, one can frequently obtain a better picture of the demographic process than would result from accepting the original data at their face value. Another series of applications of descriptive models results because a model establishes characteristic patterns--for example, for death rates by age. It can then be assumed that a population has mortality rates that follow one of these patterns, thereby limiting the range of choices. Stated differently, it has been found in most cases that there are characteristic relationships between death rates at age 25 and death rates at age 60. As soon as the investigator knows what one of those values is, the range of possible values for the second mortality rate is greatly restricted. This type of relationship is particularly useful in populations for which accurate statistics are not available. It has been possible in many cases to devise procedures for estimating a particular characteristic--such as the age-specific mortality rates--that employ indirect indices based on data that would otherwise be considered inadequate.

Causal models, which specify relationships among several factors, are particularly useful in predicting or forecasting population trends. The population trends are treated as extrapolations from an appropriate model, which sets forth a theory of population change.

One of the most widely held views about demography is that demographers are bad at making accurate projections of population trends. In the United States, until the baby boom, demographers were predicting that fertility would continue to decline, as it had since such information was first recorded in this country. And few demographers were able to predict the end of the baby boom. One of the current most hotly debated issues is whether there will be an echo boom and whether it will be caused by the large number of baby boom babies becoming parents themselves or by actual changes in the fertility of individuals or couples. However, as Nathan Keyfitz suggested more than a decade ago, despite the long series of failures and even in view of the uncertainties involved in making assumptions or predictions about the future, such predictions are necessary. The models that contain a theoretical framework for considering future change do provide some basis for the analysis of population growth and change.

Finally, the models serve as vehicles for research into the possible effects of changes in demographic determinants. In view of the impossibility of conducting meaningful experiments in human populations, this application is crucial. Altering characteristics in a model is an easy and feasible way to study the importance of each factor in the model system and to consider how much change in the final results is triggered by changes in various causal factors.

The following sections of this paper contain summaries of selected descriptive and causal models. In the last part of the paper we turn to a discussion of the application of such models to population policy, the link between nutrition and fertility, and demographic estimation from inaccurate or incomplete data.

Causal Models

We restrict our attention to two types of causal models: (1) those that take age-specific fertility and mortality rates as independent variables and consider their effect on population growth, age structure, and future change; and (2) those that consider the components of the reproductive process and examine the biological and social factors that directly determine fertility. Causal models of migration and of mortality are in very early stages of development, so we do not describe them here.

Stable-Population Theory

Basic research on so-called stable-population theory, the model that examines the effect of age-specific fertility and mortality rates on population change, goes back to 1907, when Alfred Lotka published his first article. He recognized, as others had before him, that the future size and composition of a population could be traced if one knew the age-specific fertility and mortality rates and if there was no migration either into or out of the population.

If the number of 64-year-olds is known this year, for example, and their mortality rate is known, then the number of 65-year-olds next year can be found quite easily by computing the number who die and subtracting it from the number alive at the start of the year. Similarly, the number of children born this year can be found by multiplying the number of women at a given age by the age-specific fertility rate for that age and summing over all ages. Technical details aside, this procedure is fundamentally the way population projections are prepared. One assumes that the age-specific fertility and mortality rates for each year in the future are known and calculates the numbers of people in each group who would be alive if they experienced that particular set of fertility and mortality rates.

Lotka asked the crucial question: What would happen to a population if it experienced the same age-specific fertility and mortality rates every single year? In a series of papers, he proved the proposition that constant fertility and mortality rates would lead ultimately to a population with a constant growth rate and a constant age structure, that is, the proportion of people in given age group would come to be an unchanging constant. All age groups would come to change at the same rate: at a positive growth rate if fertility outweighs mortality, at a negative growth rate if fertility falls below mortality, at an unchanged total level if these two basic forces are just in balance. Thus, a population experiencing a long history of constant fertility and mortality rates would become a stable population. The special case of a stable population with the growth rate of zero is called a stationary population. These propositions have been upheld, although Lotka's proof has been refined mathematically by later demographers.

Surprisingly enough, the populations of a great many countries do display age distributions that change only slowly. More recent research has demonstrated that if mortality has not been constant but instead has been gradually changing, the age distribution of a population will still resemble closely that in a stable population with the same fertility and mortality rates. The demographic situation in many developing countries has until recently been characterized by declines in mortality and little change in fertility. Hence, their age structures have been approximately stable. As sometimes happens to basic research, Lotka's work was treated as mathematical oddity and forgotten or ignored for many years. Not until after World War II was the potential for application of this theoretical model recognized. We return later to applications of some of these ideas.

Models of the Reproductive Process

The second major object of causal modeling in formal demography has been the reproductive process. Demographers who pioneered in this area, like Louis Henry in France and Mindel Sheps in the United States, had to be cognizant of the growing body of knowledge of the biology of human reproduction and to be aware of the social controls that serve to push reproduction to its biological limits or to control it and keep it well below any biological maximum.

It is now accepted that only a few factors are crucial. They can be grouped into those that influence the length of reproductive life and those that affect the number of births within the childbearing span.

Factors that affect the length of reproductive life include:

- age at menarche for women, which represents the lower boundary of the time when they can possibly conceive;
- age at marriage, the social control on entry into sexual relations in many societies;
- age at sexual maturity for males;
- age at widowhood or separation or death, which ends sexual relations for the surviving partner; and
- age at sterility for women and men.

In many societies, marriage patterns serve as an effective brake on fertility simply by reducing the number of years in which a woman bears children. If in two populations all women end childbearing at age 40, but in one society they marry at 15 and in the other at 30, the number of childbearing years drops from 25 for the first population to 10 for the second.

Within the reproductive years the intervals between births become the object of analysis. Again, obviously, women who have children every 2 years can have more children in a 15-year reproductive span than women who have children every 4 years. What determines the length of these intervals? After the birth of a child, a woman experiences a period during which she does not ovulate and so is temporarily incapable of conceiving again. Once she begins both ovulating and having sexual relations, even if the couple is using no contraceptive method, it may take some months before she again becomes pregnant. The pregnancy may end in a spontaneous abortion or miscarriage. If so, there is usually a brief period before she can conceive again, then a number of months until she becomes pregnant. If this pregnancy ends in a birth of a live child, then her birth interval can be broken into the nonsusceptible period following a birth, the waiting time to conception, the time taken up by the aborted pregnancy and the time until she resumed ovulating once again, another period of waiting until the next conception, and the final pregnancy itself. This cycle is shown in Figure 1.

Starting early in the 1950s, demographers in several areas of the world began to develop mathematical models

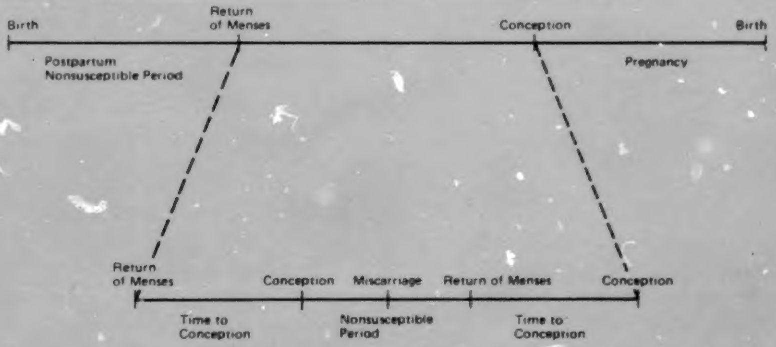
Birth
Return of Menses
Conception
Birth
Postpartum Nonsusceptible Period
Pregnancy
Return of Menses
Conception
Miscarriage
Return of Menses
Conception
Time to Conception
Nonsusceptible Period
Time to Conception

FIGURE 1 Components of the interval between two live births when there is an intervening pregnancy that ends in a miscarriage. SOURCE: Jane Menken and John Bongaarts (1977:263). Copyright © 1977 by Plenum Publishing Corp. Reprinted by permission.

that would put these components together and consider how they influenced fertility. At the time they were doing their basic work, relatively little was known about the variation in these components of fertility among populations. Their fundamental research using these models demonstrated that certain of the factors could have significant effects on fertility. If in some populations, for example, women took a long time to resume ovulating or there were strictly observed taboos on intercourse, then fertility would be low because birth intervals would be longer.

Thus, in addition to the mathematical results, the birth interval models or reproductive models provided the framework for basic biological research by indicating which factors could have significant effects on fertility and by demonstrating the need for fundamental information. Studies designed as a consequence of results from these models have demonstrated that there is extreme variation in the resumption of ovulation, from approximately six weeks on average in U.S. women who do not breastfeed to nearly 18 months in rural Bangladeshi women, who rarely wean a child before the next is born. We return to other applications of these models shortly.

Descriptive Models

Formal demography has also been concerned with the investigation of regularities in the age patterns of vital rates. On the basis of existing data, descriptive models have been devised for mortality, marriage, and fertility.

Mortality

Mortality rates follow a typical pattern in all known societies. Infant mortality is relatively high. Mortality rates decline during the early childhood years, reaching quite low levels especially between ages 5 and 15; they slowly increase thereafter, then rise more and more sharply, beginning about age 45 and extending throughout the adult years. This reversed-J pattern is found in all populations for which reasonably good data are available. Populations differ, however, in the levels of mortality (whether all rates are relatively high or relatively low) and in the relationship between infant and childhood mortality and adult mortality. The efforts to model these rates have consisted of attempts to find a mathematical formula which, given some sort of parameter representing the overall level of mortality, would yield an estimate of the mortality rates at every age. These efforts have not been particularly successful, although work continues on development of curves that fit certain parts of the age distribution.

The more useful approach to mortality models has been through analysis of empirical data. Coale and Demeny examined mortality rates by age from nearly 200 populations. They found clusters of populations within which there were differences in the level of overall mortality, but the relationship of infant and childhood mortality to mortality at older ages appeared similar. Within each of these groups they were able to relate mortality at each age to an overall measure of the level of mortality, such as the expectation of life at age 10. On the basis of their work only two pieces of information--the appropriate cluster and the overall level of mortality--are needed to identify the entire age pattern of mortality fairly well.

The Coale and Demeny tables are published for average life-spans from 20 years to 77.5 years in increments of 2.5 years. For any level between two published tables, the appropriate mortality rates can be calculated either by interpolation or by using the relational equations provided. These models have proved enormously useful in at least two ways. In projecting population growth and composition into the future, the tables offer a way of predicting mortality change. It seems likely, and has been shown to be true for a number of countries, that as mortality declines, the pattern remains within the same cluster or family but the overall level goes down. Therefore, it can reasonably be assumed that the course of mortality change follows a path described by the age-specific mortality patterns from successive levels in the model tables.

These models also play an important role in methods of determining mortality levels in developing countries, for which data are usually incomplete or inaccurate. Certain indicators of mortality levels may be obtained more easily than age-specific mortality rates. If the relationship between a mortality indicator and the level of mortality for the model tables can be found, then the age pattern of mortality can be estimated. We return to applications of this type later.

The Coale-Demeny tables have proved very useful, as already mentioned. However, there are examples of well-documented mortality patterns that lie outside the range of their models. Consequently, one area of research in formal demography has continued to be the development of better models for mortality. Only one alternative system is described briefly here, although others exist and are being developed in various demographic centers around the world.

William Brass has proposed a relational system in which the mortality rates by age from one particular population are selected as the standard. From those rates the survivorship curve--the proportion of individuals surviving from birth to each age--can be calculated. Brass has suggested that a particular relationship exists between survivorship curves in other populations and the particular one chosen as a standard. This relationship has two parameters, one that sets the overall level of mortality and the other that governs the relationship between child and adult mortality. Therefore, using the Brass system, one can describe the survivorship pattern in a population by specifying the standard to which it refers and the two parameters of the relationship. This system has also provided a mechanism for predicting how mortality will change in the future in a population with declining mortality and for devising estimation procedures.

Models of Marriage Patterns and Fertility Rates

Before going on to describe some of the formal demographic work on models for other types of population processes, we comment briefly on the emphasis we have been giving to estimation procedures. The goals of economic development programs, specific health programs, family planning programs, and many other social programs have been to improve the human condition. If a health program is succeeding, one expects mortality at various ages to decline. It is therefore crucial to be able to determine the level of mortality at various time periods with some degree of accuracy in order to be able to assess whether change has actually occurred and the magnitude of any change that has taken place. Likewise, in historical studies, if one wants to elucidate the causative factors that have influenced change in the past, it is essential to be able to measure whether change has occurred and its magnitude. In many cases the kinds of detailed data required for the standard kinds of mortality statistics we expect from developed countries today simply are not available. The development of measurement techniques that make it possible to use the less detailed data or fragmentary information that may be available can be likened to the development in natural sciences of methods or instruments that made possible the measurements necessary for important advances. The Nobel prize in medicine has been awarded several times to individuals whose contribution was the development of methods or instruments rather than the discovery of new facts. It has been awarded, for example, for the development of tissue culture techniques, which allowed the polio virus to be grown and subsequently led to the development of effective vaccines. For the study of population and of the determinants and consequences of population change, the measurement methods devised as a part of formal demography or an application of formal demography may be equally important.

Models have been developed for both marriage rates by age and for fertility by age within marriage by examination of age patterns from a sufficient number of populations to be able to discern commonalities. Working with data from developed and developing countries, Coale noted that the patterns of marriage rates by age are similar. These patterns differ primarily in the proportion of the population who ever married, the lowest age at which a minimum significant number of marriages took place, and the general speed of entry into marriage. He was therefore able to describe marriage distributions in a wide range of populations by a single curve with three parameters. It is not as easy as it is for mortality to ascribe biological reasons to the finding of regular marriage patterns, except in relation to sexual maturation. However, maturation alone could not explain the variation in the observed nuptiality patterns. Coale has suggested that there are behavioral and cultural characteristics that make it reasonable to expect that a common curve would describe marriage patterns fairly well.

For fertility within marriage, Coale and Trussell, extending earlier work by Louis Henry, found that age patterns of fertility within marriage could be described by an empirically based curve that had two parameters: One represents the overall level of fertility; the other represents the extent of fertility control within marriage. They were then able to combine the curve for marriage patterns and the curve for fertility within marriage into a description of fertility rates by age of the woman.

Although derived under the assumption that no fertility occurs outside marriage, the new models (with appropriate changes in parameter values) have been found to give tolerable descriptions of overall fertility even when there is a significant amount of nonmarital childbearing.

Other descriptive models have been proposed for fertility. They all work to serve the goal of providing adequate representations of fertility patterns by age and allowing projections of future fertility to be made by assuming a particular general subsequent course of fertility (i.e., specifying changes in the pattern and level of the age-specific rates over time). They also provide a basis for measurement techniques and facilitate comparative studies because they summarize the full set of fertility rates into only a few parameters for comparison among different populations.

The Coale-Trussell fertility models have also proved useful in detecting the use of fertility control in populations in which control is exerted primarily by terminating fertility after the woman reaches a certain age or after the desired family size is achieved, in contrast to populations in which methods of fertility control are used for spacing purposes. Knodel has estimated the fertility control parameter of their model for a number of Asian countries over time. The results give plausible estimates of the time at which fertility control began and the rapidity of its adoption by larger and larger segments of the population. The information necessary for deriving these estimates consists only of age-specific (marital) fertility rates. No direct information was collected on contraceptive use or other methods of fertility control. Thus, in this case the models provide a glimpse of processes that are unobservable, since no direct data were collected at the time this type of change was taking place.

APPLICATIONS TO POPULATION POLICY

Age Structure

Population and economic policy in many countries is particularly sensitive to the age structure. When there are relatively large numbers of children, as in many developing countries, and especially when these numbers are increasing, the need for education facilities is acute. When the number entering the labor force is rapidly increasing, the economy must expand rapidly to provide additional jobs, or unemployment and underemployment will increase. In developed countries the age structure poses wholly different problems, because of large and growing proportions of old people. The services required by an aging population are quite different from those needed by young children.

The proportion of the population of working age is an important demographic indicator of the productive capacity of the population. The dependency ratio is one way of indicating how many people must be cared for by the work of those in the productive ages. This measure is usually defined as the number of people under age 15 plus the number 65 and over divided by the number in the age group 15-64. Unless disaggregated, however, this ratio does not indicate how the burden is distributed between old and young.

What then determines the age structure of a population? This question has been approached through the theory developed in formal demography of stable populations. The age distributions of Sweden and Mexico are shown in Figure 2. Sweden is the classical example of an "old" population; a high proportion is in the older age groups. Mexico, on the other hand, has a very young population. In 1970, for example, 46.4 percent of the population was under age 15 in Mexico, compared with only 20.9 percent in Sweden. The usual reaction on examining these graphs is that Sweden is an old population because it has such low mortality. Indeed, the expectation of life in Sweden is over 70 years for both men and women, whereas Mexico has much higher mortality. However, as mentioned, the work on stable population theory has demonstrated that mortality has relatively little impact on the age structure of a population. If the mortality rates were exchanged--if Mexico had Sweden's mortality rate and vice versa--the age pyramids we see would be relatively unchanged. A population ages because its fertility declines, not primarily because survival increases. All countries with a relatively long history of high fertility rates exhibit the typical pattern of a steep slope of population with age. Declines in mortality increase the growth rate of a population, that is, increase its overall size, but they have relatively little effect on its age structure. In fact, counter to common intuition, declining mortality makes the population (slightly) younger, not older, since the greatest relative declines in mortality tend to occur in infancy and childhood. A decline in fer-

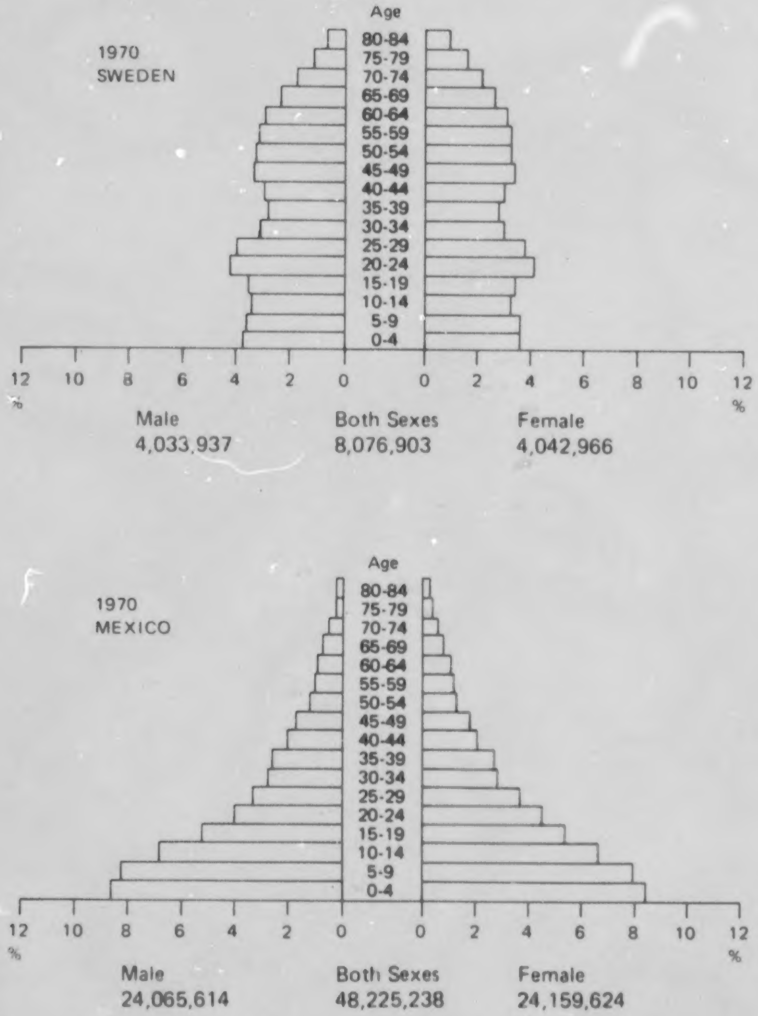

FIGURE 2 Population by age and sex in Sweden and Mexico, 1970. SOURCE: Institute of Developing Economies (1970: 119 and 49).

tility does, overall, reduce the growth rate and thereby lead to a population with higher proportions in the older age groups.

Changes at the individual level in fertility do not translate directly into similar changes in the overall birth rate. Understanding of this phenomenon has come about through examination of the effect of age structure on future population growth, again through the use of formal demographic methods. One measure of fertility frequently quoted is the total fertility rate, the average number of children a woman would bear if she were subject to a specified set of age-specific fertility rates. Total fertility rates can be calculated either from the age-specific fertility rates in a given year (usually known as a period total fertility rate) or from the actual rates experienced by a group of women as they pass through their life-span (cohort total fertility rate). From stable population theory, we know that a low-natality population is stationary, i.e., nongrowing, if over long periods of time its fertility rates have been constant and are such that the total fertility rate is about 2.1. Intuitively this result is understandable if one thinks that a woman who lives through the reproductive span has to have one child to replace herself, one to replace a male partner, and a small extra fraction to replace those woman who died before completing the reproductive span. If a population has a total fertility rate below this figure for a long period of time, it will begin to decline. The total fertility rate is sometimes expressed as the sum of the age-specific fertility rates for having female children. In this case, to allow for the females who die before completing the childbearing years, the replacement level of fertility would be slightly over 1.0. For most of the 1970s in the United States the total fertility rate calculated in this way was well below 1--as low as 0.85. According to the current age-specific fertility rates, 100 women would replace themselves with about 85 daughters. It would seem logical that the population would decline. Ultimately this statement is true; however, the results of research in formal demography on age structure indicate that there is a built-in momentum of population growth in a population in which fertility is declining. If fertility has been relatively high, then relatively high proportions of women are in the childbearing ages or will enter the childbearing ages in subsequent years. The ultimate decline will not begin to take place until this bulge passes through the childbearing years.

In a population that has experienced a recent decline from high levels of fertility, the bulk of the population is in age groups with low mortality. Therefore, the population death rate is low for two reasons: Since fertility has declined, fewer infants are born. This change alone serves to reduce the death rate because infants are at especially high risk of mortality. The number of deaths in the older age groups is not yet particularly high. Therefore, because so many in the population are women of childbearing age, even if they are not having enough children to replace themselves, they are having enough children over the short term to more than replace the people who are dying each year. Not until the babies born in the lower fertility years move into the childbearing ages would the growth rate become negative.

Age Structure of the United States

The results of changes in fertility can be seen in the U.S. age distribution for 1980, shown in Figure 3. This is a distinctly unusual age distribution, reflecting major changes in fertility that have occurred over the last 50 years. In the older age groups, from age 60 on, there is the typical steep decline that reflects higher fertility at the start of the century and, to some extent, the high mortality rates of older people. From 1900 to the 1930s when relatively few children were born, fertility fell. The most distinctive feature of the U.S. age distribution is the "baby boom," which began after World War II and peaked in 1957. This tidal wave of people has been moving through the age structure of the United States, experiencing special problems and encountering special circumstances attributable to a great extent merely to the enormous size of their peer group: Schools unprepared to accept such a large influx of young people and a labor market forced to adjust to the entry of huge numbers of young and inexperienced workers are only two of many examples.

Subsequently there has been the so-called baby bust of the 1970s. The baby boom cohorts are now in the prime childbearing ages. They are bearing children at below replacement levels in the sense that each woman, if current fertility rates continue, will have, on average, fewer than the 2.1 children necessary for replacement. However the overall growth rate in the United States is still positive, because there are large numbers of people

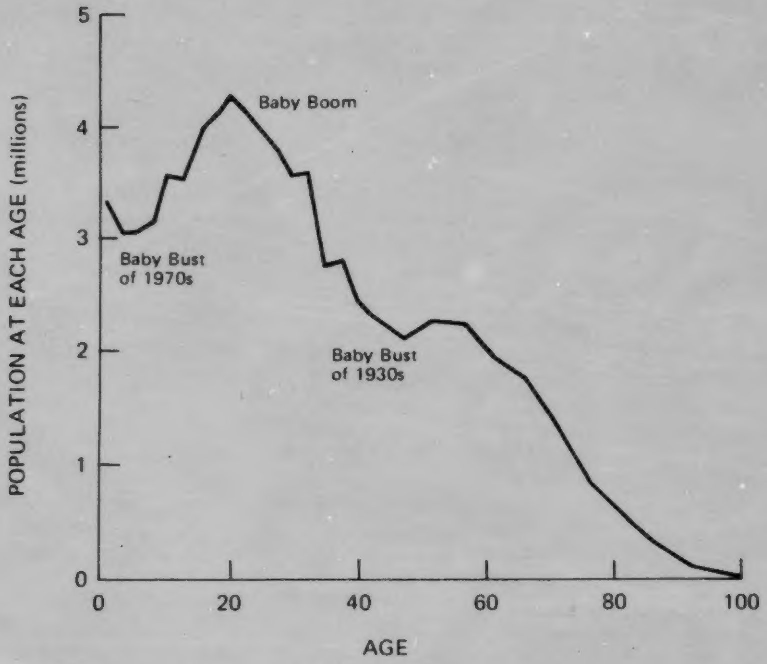

FIGURE 3 Age distribution of the population of the United States, 1980. SOURCE: U.S. Office of Management and Budget (1981:49).

in the childbearing ages. Using the techniques of population projection derived from the studies that led to stable population theory, it is possible to calculate when the growth rate would become negative. The change would not take place until close to or soon after the year 2000.

This unusual age structure has predictable implications for U.S. society over the short and long term. The outlook for the future is now being taken into account by U.S. government policymakers, as exemplified by the inclusion of discussions in the federal budgets for 1980 and 1981 on the impact of the baby boom generation on future needs for education, medical care, productivity of the labor force, housing, retirement and health, and long-term care of the elderly.

It should be pointed out that the projections of the U.S. population into the future are probably quite accurate when we are considering the survivors of people already born. Therefore, over a projection horizon of up to 30 years, we can predict the numbers of people in the age groups above 30 with reasonable accuracy. It is highly unlikely that age-specific mortality rates will rise significantly; the major uncertainties come from the assumptions we make about future declines in mortality and the numbers of immigrants added to the population. There can be some surprises, however.

In recent years, mortality among the oldest Americans--those over 70--has been declining more rapidly than assumed in most projections. Therefore, the population of older Americans is increasing more rapidly than predicted. However, this error is far smaller than that expected in the predictions of numbers of births in the future. In a society in which the use of methods of fertility control is nearly universal, fertility rates can fluctuate rapidly and can have rather large swings, as can be seen in the history of the U.S. population over the past 30 years.

The baby boom took demographers by surprise. They saw no reason to expect that the long-term moderate decline in fertility in the United States would not continue. The baby bust of the 1960s and 1970s also was not predicted. The modest upturn of births in recent years was correctly predicted. It was anticipated because the baby boom generation is so large that, despite its low individual fertility, it was expected to contribute a large number of infants to the next generation. There remains an enormous controversy over whether there will be another baby boom, caused by an increase in fertility rates.

Evaluating the Impact of Policy Alternatives

The concepts and techniques of formal demography have also proved useful in projecting the future course of population change in developing countries. They have been useful, too, in evaluating the impact of policy alternatives. Many countries of the developing world are characterized by very early and nearly universal marriage for women and rather high fertility within marriage. Lesthaeghe employed models of marriage and marital fertility to assess the impact of changes in marriage patterns. Selecting marriage models similar to patterns in early-marrying countries, he projected populations characterized by young ages at marriage. He assumed that changes in fertility within marriage that are characteristic of increased use of contraception would take place. He then produced population projections incorporating successively higher degrees of fertility control. He found that, unless fertility control exceeded that in the developed countries with the lowest levels of fertility, these countries could not end population growth while their current marriage patterns continue. No amount of effort put into reducing the fertility of married women could bring the growth rate down to zero.

Raising the age of marriage, on the other hand, could have a major impact. Moderate increases in age of marriage, combined with moderate use of fertility control methods, would allow relatively rapid decline in overall birth rates in the population. The policy implications are clear. If population growth is to be reduced, a two-pronged program designed to postpone marriage and to reduce fertility within marriage is needed.

Major declines in fertility in a number of Asian countries (for example, Korea, Singapore, and China) have taken place in recent years. A great portion of this decline is attributable to increases in the age at which women marry. In some countries this change may have been accomplished in part through laws raising the minimum age at marriage, as in China, as well as through societal changes that have served indirectly to increase the age at marriage (for example, major increases in the educational attainment of women). Again, the demographic models allow us to examine the implications for the future of continuing or changing patterns of demographic events. The value of this perspective for evaluating alternative policies seems to be accepted by many of those responsible for policy decisions.

However, the demographic implications of policy changes are sometimes not considered in advance. The classic example is Romania in 1967. Fertility had been at low levels for a number of years, primarily because of the use of abortion to curtail drastically pregnancies. Concerned about future declines in the labor force and the effect of repeated abortions on the health of women, the Romanian government decided to terminate the availability of abortion. The birth rate climbed precipitously six months later, as those women who would have obtained abortions bore babies instead. Slowly, over the next months and years, the birth rate declined as women found other means of preventing birth. But 10 years after the major policy change, the Romanian birth rate remained about a third higher than it was in the early 1960s.

Obviously the policy change was successful in achieving the goal of increasing fertility. However, Romania must live with the consequences of the abruptness of the change. The cohort born in the year after the change is nearly three times as large as the next older cohort. It is also followed by much smaller cohorts. What does a country do to adjust to this huge spike tracing its way through the age distribution? Maternity wards, then schools, were swamped; members of the cohort will also encounter difficulty in entering the labor force and in finding housing.

China

In China today, policies are being adopted that, while addressing immediate problems, may create formidable problems for the future. China has reduced its birth rate through a combination of increasing the minimum age at marriage and instituting powerful disincentives to high fertility. Recently here has been much emphasis on reducing population growth by encouraging one child per family. The emphasis on such a severe reduction in family size is related to a target of reducing the rate of population increase to zero by the end of the century. Demographic analysis brings to light the difficulties of the widespread adoption of a policy of one-child families and the possibly undesirable consequences of the low fertility that would be needed to meet the target of zero growth in the year 2000. The most obvious effect is a massive change of family structure. Children would grow up without siblings. A generation later, aunts, uncles, cousins would all be practically nonexistent for any given child. Either families would become lineal structures, or new types of "families," composed of individuals not related through close blood ties, would have to be established. Rules for army service, for example, would have to change. Currently, men who have no brothers are excluded from military service; if everyone is an only son, no one would be eligible.

The principle of population momentum described earlier may make it impractical and undesirable to retain the zero-rate target officially suggested in China. Because of high birth rates in the recent past, the number of persons of parental age will peak in the year 2000, no matter what changes in fertility are introduced in the interim. In that year women would in fact have to bear children at a rate lower than ever observed in any national population to attain the zero-growth target. The total fertility rate would have to be temporarily far less than 1.0. Moreover, the consequences of attaining the stated target may themselves be undesirable 30 or 40 years into the next century. The large numbers born in the 1960s and early 1970s would be approaching old age. The result would be an extremely distorted age distribution, in which there would be more than twice as many persons in their 60s than in any 10-year span under age 40. Alternative population projections show that a policy aiming at a reduction in fertility below the long-range replacement level but not to the extremely low levels envisaged, combined with a gradual return to the replacement level, may be more practical to attempt. It would yield a population somewhat larger and attain a zero rate of increase a little later, but with a less irregular and perhaps less problematic age distribution.

Fertility Policy

Models of human reproduction have also served as vehicles for evaluating alternative policies and for elucidating potentially important determinants of fertility. Among the more fruitful areas of research has been an examination of the effects of changing components of birth intervals. For example, when family planning programs were being adopted in many countries, it was assumed in a number of instances that a reduction in the birth rate would be directly related to the increase in the proportion of women adopting use of an effective contraceptive. Some perhaps overenthusiastic planners thought that if 20 percent of women began using contraception, a 20 percent reduction in the birth rate would follow. When a reduction of that magnitude did not take place, it was assumed that family planning programs had failed. One reason for the failure of the anticipated reduction is that women who initially use family planning services are typically those who already have relatively low fertility, often because they are older. Demographic models, however, allow us to appreciate a fatal flaw in this type of reasoning, even when all women experience equal fertility rates, as is approximately true when attention is confined to a specific age group of women with specific socioeconomic characteristics. Recall that we said in the discussion of reproductive models that the rate of childbearing among married women was inversely related to the length of their birth intervals. If the birth interval is two years for all women, the birth rate among married women will be 500 per 1,000 women per year. The birth rate will decrease by 20 percent if the birth interval increases by 20 percent. Contraceptive use, however, affects only part of the birth interval: the period between the resumption of both ovulation and sexual relations following a previous birth (or their initiation following marriage) and the time of conception. If the contraceptive chosen is perfectly effective, so that the 20 percent who adopt it have no further births, the birth rate is indeed reduced by 20 percent also. This result implies that the remaining 80 percent of women continue to have a child every other year, so that the birth rate of all women is 0.80(1/2) = 0.40 or 400 per 1,000 women per year instead of the original 500.

Unfortunately, no methods of fertility control except perhaps sterilization can achieve perfect results. There are contraceptive failures; moreover, individuals discontinue use for a variety of reasons, ranging from discomfort from side effects to lack of availability of supplies to deciding to have another child. Therefore, for the 20 percent who adopt use, the time to conception is prolonged, but all births are not eliminated. A contraceptive that is 90 percent effective will reduce the monthly conception rate to a tenth of its original value and increase the time to conception by a factor of 10, according to results from formal demography. Returning to our original case of a birth interval of two years, assuming that the time to conception was five months and there were no spontaneous or induced abortions, what would adoption of a 90 percent effective contraceptive by 20 percent of women imply? For the 20 percent who use contraception, the time to conception would be 50 months instead of 5 and the birth interval would increase by 45 months, to 5.75 years instead of 2 years. For all women the birth rate would be 0.8(1/2) + 0.2(1/5.75) = 0.4 + 0.035 = 0.435, or 435 per 1,000 women per year. The reduction is from 500 to 435 or 13 percent instead of the expected 20 percent. Since we have used a very high effectiveness rate unlikely to be realized in practice in a group of individuals just beginning contraceptive use, this figure is almost certainly unrealistically high.

Another reason the estimated reduction is too great is that research has demonstrated that women who begin using contraception tend to be older than the average woman of childbearing age. Since fecundity declines with age, the birth intervals of new contraceptors tend to be longer than the average for all women. Again, use of demographic models can help assess realistically the potential of policy alternatives.

Another similar example relates to the effect of induced abortion. One abortion prevents one birth. But it does not prevent one birth in the sense of reducing by one the children ever born over a woman's lifetime. Again, reference to formal demography can elucidate this apparent paradox. An abortion increases the birth interval in which it occurs by only a few months--the time for the pregnancy and short postpartum period (usually only two or three months) and the time to conceive again. Thus, it removes only a relatively small number of months from the reproductive span. An abortion would have to delete entirely an average birth interval to reduce the lifetime number of births by one. The conclusion is inescapable: Abortion alone is not an effective means of fertility control except when used repeatedly, perhaps with consequent ill effects on the woman and requiring the costly provision of abortion services. When accompanied by even moderately effective contraceptive use, however, abortion can reduce fertility significantly; this strategy in essence makes the contraceptive 100 percent effective.

NUTRITION AND FERTILITY

The development of models of the reproductive process established a framework for analysis of other influences on fertility. Several years ago it was proposed that nutrition was closely related to fertility. A plausible argument was made that the fertility of women in many developing countries was well below the biological maximum because they were poorly nourished. It was well known that, under famine conditions, birth rates drop. Could it be that women who were chronically malnourished but not famished also suffered from depressed fertility due to their nutritional condition? The policy implications of this hypothesis, if it were true, were considerable. Programs designed to improve nutrition, to provide more food to many parts of the world, might have the unintended effect of increasing fertility and increasing the number of people who required food, thereby defeating their original purpose.

Five years ago little information existed to judge whether this hypothesis was indeed true. But fragmentary evidence from a variety of sources suggested that it was a plausible hypothesis, one that required attention. In the interim, carefully designed studies have examined it. The existence of an explicit framework for the analysis of fertility was extremely important. The human reproductive process was divided into either biological or socially determined components, and the effects of each of these components on overall fertility were analyzed. If nutrition affected fertility, it had to affect some or all of its components, either by prolonging the time to resumption of ovulation, or increasing the rate of spontaneous abortions, or increasing the time to conception, or postponing menarche, or leading to earlier sterility. Studies were designed to concentrate on the effects of nutrition on each of these components. The balance of the data does not support the proposition that nutrition significantly affects fertility, except when malnutrition approaches the extremes of famine. Therefore, although the relationship between nutrition and fertility is not completely elucidated, it now appears likely that improving the nutritional status of mothers will not have the inadvertent effect of increasing their fertility.

ESTIMATION FROM INCOMPLETE DATA

In most developing countries, birth and death registration systems are frequently absent altogether or function poorly. One cannot reliably obtain even the simplest demographic indices, such as birth and death rates, from central statistical offices. In such situations the techniques of formal demography have proved invaluable. In some instances, estimates based on models provide the only clue about the demography of a population. In other cases the demographic methods derived from formal demography provide measures of the completeness of recording in registration systems.

An early use of models in this spirit followed the development of model stable populations. By reversing the logic of the Lotka theory, demographers found they could provide fairly good estimates of fertility and mortality from the age distribution obtainable from the census. Many of the baseline demographic estimates in India, for example, were obtained by using model stable populations and newer techniques. Newer developments in methodology have moved away from reliance on the assumption of stability (fairly constant fertility and mortality in the recent past), since it is less and less likely to hold.

Descriptions of two techniques provide some idea of their power and use. The first technique is used to estimate fertility. Often two questions, one on births in the last year and another on the number of children ever born, are asked of women in large surveys or censuses. When the results are tabulated (by age of the women), it is often apparent that there are errors in answers to both questions. Frequently, the average number of children ever born, when tabulated by age, displays the most obvious error; the average falls at the higher ages. Women who are 45-49 report fewer children ever born than women 35-39. Taken at face value, these figures suggest a rise in fertility. This finding is observed even in populations in which it is fairly certain that no fertility change has occurred. Careful analysis has shown that women tend to omit children who are no longer in the home, and older women are more likely to have children old enough to have left home. Numbers of births in the last year, on the other hand, may be understated or overstated, but the former occurs more frequently. The reason seems to be an error in the reference period on the part of the respondent. She may, on average, report births in the last 10 rather than 12 months. Methods have been devised, however, to exploit to the fullest the information contained in answers to both these questions. For example, in one widely applied method, the age pattern of fertility is obtained from the question on births in the last year. The assumption is made that the reference period error is on average the same for all women, regardless of their age. The level of fertility is obtained from responses to the question on children ever born as given by younger women, whose answers tend to be most accurate because all of their children are young enough to still be living with them. A calculation can then be made of the number of children women 25-29 should have ever had if the age-specific fertility rates held. This expected number is compared with the reported number. If women report more births than expected, then the reference period was too short; the information on births within the last year was really information on births in a shorter period so that the level of fertility would be underestimated. The difference between the observed and the expected can be used to adjust the rates and inflate them upward to match the reported actual numbers of children ever born. If the expected values are higher than the observed ones, then the reference period was too long and the fertility rates can be adjusted downward. Thus, using information that most women can provide even in countries where statistical data systems are poor, formal demography often provides a means of obtaining reasonably accurate information on levels of fertility.

The second technique is used for estimating childhood mortality. From two questions asked of women in a census or survey, one on the number of children ever born and the other on the number of children still surviving, it is possible to compute the proportion of children who have died, specific to age groups of women. These proportions can be translated, using techniques of formal demography, into standard life-table measures of the proportion of infants born who died before their first, second, or fifth birthday.

Together these two techniques have provided the best available estimates (and in many cases, the only ones) of fertility and mortality in large parts of Asia, Latin America, and Africa. Their importance cannot be overstated. Later research has resulted in the development of other techniques similar to these two. The demographer now has a large tool kit of diagnostic aids and techniques for providing useful estimates of fertility and childhood and adult mortality. The problems have by no means all been solved, however. There is much room for improvement of our techniques for measuring change in fertility, especially when it is dropping rapidly.

Problems involving incomplete or inaccurate data are not confined solely to developing countries. One example in the United States serves as a reminder. Techniques of formal demography have provided the basis for estimating the completeness of enumeration in the 1950, 1960, and 1970 censuses. These estimates, which show that about 2.5 percent of Americans were not counted in 1970, have provided the basis for the demand to adjust the 1980 census.

FINAL REMARKS

The techniques and results of formal demography have proved their value in practical situations, yielding information based on new measurement techniques, providing the analytic framework for evaluating problem areas, and serving as a vehicle for projections of consequences of alternative policies. Much remains to be done. Applications of existing results from basic research surely have not been exhausted; however, only continuing emphasis on basic research in formal demography holds the promise of innovative new results that would open up whole new fields of applications to the social problems facing the world.

REFERENCES

Institute of Developing Economies
1976 Age Pyramids of the World Population 1950-1970. Tokyo: Asian Economic Press Ltd.
Menken, Jane, and Bongaarts, John
1977 "Reproductive models and the study of nutrition--fertility interrelationships." In Henry Mosely, ed., Nutrition and Human Reproduction. New York: Plenum Press.
U.S. Office of Management and Budget
1981 Budget of the United States Government. Fiscal Year 1981. Washington, D.C. U.S. Government Printing Office.

The Study of Voting

Philip E. Converse, Heinz Eulau, and Warren E. Miller

INTRODUCTION

Significance of Elections in the United States

America's representative democracy would be inconceivable without elections. By enabling citizens to participate in the choice of their representatives, elections are the only events in which the American people, however bounded--by nation, state, or locality--experiences its common existence and directs its common fate. This happens in the perhaps primitive awareness that, in the privacy of the voting booth, one person's vote is equal to another's and that from this circumstance emerges a legitimate, collective decision as consequential for the individual as for the community. Elections make it possible for a people to express its support for or opposition to the government and its policies as well as to consider and vote for alternatives. There are in democratic elections winners and losers among both electors and candidates for office, but whatever the outcome it is the electoral act that confirms the viability of the political system in which it can be freely performed.

This stylized brief may appear merely to articulate a political myth; we believe it describes the American reality. As one views the American electoral system from the perspective of the country's own history as well as in comparison with most other nations, it has proved to be extraordinarily stable yet adaptive because it is continually reformed in the face of new conditions and exigencies. Its many difficulties and shortcomings notwithstanding, the American electoral system is one of the great institutional wonders of the political world, especially at the national level. Periodically and regularly, tens of millions of citizens go to the polls on a single day to cast secret ballots and, through the simple act of voting, to give direction to what is properly called popular government. Changes in the composition of representative offices and changes in public policies occurring as a result of voting may not be many, rapid, or great, but they do have some important short-term and many critical long-term consequences for the circulation of governing elites, the representation of diverse interests, and the evolution of public policies.

Involved in voting and elections is a fundamental issue, perhaps the most fundamental issue, of any political system--the relationship between rulers and ruled, governors and governed, leaders and followers, representatives and constituents. If, in a precise formulation, politics is concerned with "who gets what, when, how, and with what effects," elections in the United States are the crucial social mechanisms that define the relationship and establish the linkage between the people and its government. Elections in the American democracy have as a primary function the final choice of the people's representatives. The primary function of representatives, in turn, is to make collective decisions on behalf of the represented as well as to serve them in various other ways. And in order to govern effectively in the American democracy, representatives and would-be representatives must seek their constituents' confidence and support by winning elections. There is no other way.

Significance of Voting Behavior Research

Voting behavior research derives its significance from the importance of elections in American governance. Yet theories of democracy make demands on the individual voter, as do models of rational choice, or on the electoral system, as do models of pure competition, that no human actor can truly live up to or that no social mechanism can easily satisfy. Because the premises of democratic theory and the promises of the democratic faith are often disappointed by actualities, there is a tendency on the part of the public to doubt the societal importance of democratic elections; and there is a tendency among some scholars to underestimate the careful scientific study of the apparently "simple act of voting" and the apparently simple procedure of vote aggregation that constitutes an election.

In particular, it is sometimes charged that the scientific study of voting behavior does not address the really significant problems of American society--the issues of war and peace, of wealth and poverty, of power and liberty. These are admittedly unsolved problems of American democracy. But because these problems exist and still seem to elude scientific understanding and political reconstruction, it does not follow that investment of scholarly time and effort or investment of societal resources in conducting scientific research on voting behavior is wasted on something socially and politically insignificant. As is clear throughout this paper, the simple act of voting is not simple, and the modern American electoral system in all of its aspects is enormously complex. Any preconceived judgment as to the social significance of scientific work precludes what is yet to be concluded--in this case the role of the citizen as voter and the function of elections as mechanisms of public choice and political control in the American democracy.

Basic research on voting is "basic" because, unlike historical descriptions of the specific flow of events in particular election situations, it aims at a much higher level of generalization that can subsume particular elections as intelligible special cases. Its initial questions are basic ones as to how the citizen's mind is made up: first, whether to participate at all, and, if so, what alternative to choose. Only if these fundamental questions are answered can one address the much more difficult but equally basic questions of how individual votes are aggregated to make for a collective decision and why collective decisions are what they are. These questions sound simple, yet scientifically valid and reliable answers are difficult to come by. Electoral research, originally dependent on historical documentation and aggregate voting statistics and only in the last four decades based on data collected by means of surveys, is a very young field of social science investigation. Much has been learned, but as with any area of serious inquiry, each good answer brings into focus an array of new questions.

Voting research using aggregate statistics, and later survey interviews, began with a few of the simplest, most accessible demographic indicators to explain the behavior of voter groups and election outcomes: income, education, social class, urban or rural residence, race, religion, and so on. Today the number of variables used to illuminate the simple act of voting and its collective significance is in the hundreds: from personality-rooted motivations, learned social and political attitudes, role-dependent perceptions and expectations, interpersonal relations and pressures, group-linked and institutional constraints, as well as many other such psychological or sociological variables, to the effects of the mass media of communication, the self-presentations and efforts of candidates in campaigns, the impact of long-term social or economic conditions as well as of unexpected events, the influence of campaign finances, the activities of party organizations and interest groups, the facilities or impediments of the electoral mechanism itself, and so on.

Increasingly, electoral research, especially at the subnational level, has come to focus on the transactions--cognitive and interpersonal--that take place between representatives and represented, not only in the campaign and election periods but also in the interelection period. Electoral research thus becomes interested in the emergence, ambitions, and recruitment of candidates, in representatives' "home styles" while in office, the services they perform for their constituents and the benefits they obtain for their districts, their voting records, and so on. These concerns arise from the realization that the simple act of voting is in part an extension of a much more pervasive set of relationships and interactions that link citizens and those who are elected to act in the interest of the represented. The study of voting behavior, elections, and representation has thus become a single field of scientific specialization.

Practical Concerns and Basic Research

Although a distinct and growing corpus of knowledge exists that may be rightfully referred to as basic research on voting, there is also, as with other realms of inquiry, a cognate area of engineering or practical applications that has a life of its own, although it necessarily interacts with its basic research counterpart. The two centers of such practical concerns are the government and the growing industry of commercial polling in the private sector, although the mass media in the private sector also have a considerable applied interest.

The nature of these interests is evident in each case. It is the responsibility of governments to design and run the machinery of democratic elections in ways that maximize the fidelity with which the public voice at the polls is registered. Decisions to fine-tune the machinery in new ways through electoral reforms are ultimately political ones, but they do rest increasingly on empirical information about flaws in performance. Similarly, when electoral reforms are implemented, basic research takes a keen interest in assessing their consequences, intended or otherwise.

Commercial polling agencies do a great deal of work that has nothing to do with politics. Nonetheless, some of their most visible activities include monitoring public opinion on political issues of the day and assessing trends in candidate preferences leading up to elections. In general, commercial polling activities can be divided into two categories. The first involves proprietary research purchased by campaign organizations seeking confidential information about trends in public preferences about issues and candidates. Such information can guide campaign strategies and thereby provide an electoral advantage to the purchaser. The second involves polling work designed to be broadcast to the nation as rapidly as possible; of course, this type of activity blurs immediately with the concerns of the mass media in purveying "news" about shifts in public opinion generally or assessments of political leaders and election candidates in particular in order to provide interpretations of the meaning of political events that are increasingly accurate and timely.

Both the polling and the media excursions into public opinion and electoral competition as measured through sample surveys have drawn at many points on basic research in these areas. And repeatedly over recent decades questions have been raised about the effects of these excursions on the nature and quality of public participation at the polls in actual elections. These are questions that ultimately are left to basic research to answer.

Thus the relationship between basic research into voting processes and concerns about practical applications is fundamentally a symbiotic one, with stimulation and nourishment flowing in both directions. This two-way traffic between practical concerns and basic research considerably antedates the initial burgeoning of more than casual election polling in the 1930s. An earlier speculative and literary research tradition had, for example, produced advocacies for a variety of electoral reforms around the turn of this century and had inquired with the research tools of the day as to their impact. One of the first interview-anchored studies of voting behavior

(Merriam and Gosnell, 1924) was stimulated by the puzzle that the extension of the franchise to women after World War I was followed by a decline in voter turnout, due apparently not only to the inexperience of women in voting but also to other conditions in the electorate that caused increasing male abstention.

The advent of the Gallup poll in the mid-1930s stemmed from increasing interest in problems of population sampling, then still an art but in the process of becoming a science. Its rapid ascendance into high public visibility, along with an increasing number of imitators, was important in turn in stimulating academic fascination with more basic research uses to which the new tools might be put. The first theoretically enlightening voting study (Lazarsfeld et al., 1944) profited from this stimulation.

Basic research into voting after World War II did not begin as a response to any felt malfunctioning of the electoral system or demand for reforms, but rather as a response to the fiasco of the commercial public opinion polls in misreading, for a variety of reasons, the pulse of the electorate during the 1948 presidential election (Mosteller et al., 1949). The first major nationwide, university-based study of voting (Campbell et al., 1954) was a response to the 1948 difficulties and led to an institutionalization of basic research on national elections.

Purpose of this Paper

This paper addresses the interface between basic research on voting processes and those practical concerns and applications representing some of the short-term social utility that such research may claim and, without doubt, some of its potential disutilities as well. As our compass is limited, we do little justice to either basic research or to the realm of practical applications as currently developed. However, we do start with a capsule review of the evolution of conceptual frameworks informing basic research into voting as they have developed since the sample survey began to be exploited. We then return to an examination of the more recent interchanges between basic research and more practical concerns.

THE DEVELOPMENT OF CONCEPTUAL FRAMEWORKS FOR VOTING RESEARCH

One indicator of the intrinsic complexity of the simple act of voting is suggested by the great variety of disciplinary perspectives that have engaged in intellectual traffic of either import or export direction with the general topic. At one time or another, efforts at serious contributions to basic theorizing about voting processes have been offered by mathematicians; anthropologists; economists; journalists; sociologists; statisticians; psychologists of clinical, experimental, and social bent; geographers; and historians, as well as political scientists. At the same time, processes surrounding popular voting have been taken as an interesting testing ground for more generic phenomena regarded as basic within one discipline or another, including perception, learning, attitude formation and change, cost-benefit calculi in decision making, interpersonal influence, group conflict, media impact, coalition formation, intergenerational transmission of values, and cycles of historical change, to name but a few.

Such a profusion of viewpoints can be sustained in part because voting processes can be viewed at many levels and in part because voting data are among the most accessible registers of large-scale human behavior monitored in numerous countries over relatively long periods of recent history. The use of sample survey data to decipher the evaluation processes carried out by individual voters, a "micro" topic of particular interest to psychologists, is only one facet of the subject matter. Another facet, which appeals to those interested in the logical implications of axiomatic systems, either takes the votes of individuals as given or disposes of them with a simplistic assumption or two, then focuses on what can be expected to happen under varying specified circumstances as these individual votes are aggregated into a collective social choice. Also at a macro level but in a more empirical mode are the efforts of those scholars interested in making sense of long-term historical voting records, which lack the underpinnings of contemporary sample survey data. Proceeding with vote distributions as time series and capitalizing on geographic differentiation for further diagnostic help, these students work in a fashion more closely analogous to the geologist than do either the psychologist or the logician.

From the point of view of these varying disciplinary perspectives, the chief value of voting research is less to understand voting processes per se than to use them to illuminate other phenomena seen as basic to the discipline involved. Without wishing in the least to gainsay this type of intellectual utility, it is our intention to set it completely aside and concentrate instead on the evolution of those conceptual frameworks dedicated specifically to the study of voting itself, both as individual behavior and as a key element in broader (democratic) system performance. Even within this restricted compass there is more than enough heterogeneity to keep us occupied, and we shall be obliged to use very broad brushstrokes at that.

It is most useful to begin our discussion of this evolution with the commonsense understanding of what a voting system is all about. A voting system exists as a means of eliciting preferences felt by system members concerning current political alternatives. The alternatives at stake may be competing policies, parties, or potential leaders, and the system itself may reflect a greater or lesser explicit concern with the problems of providing legitimation for controversial but binding decisions. Similarly, rules of the game defining what constitutes a "winning" vote and what outcomes victory entrains can vary markedly. What is constant, however, is the notion of votes as expressions of preference. If a single alternative is offered, then a vote is cast as simple ratification, and rejection must be expressed by some form of abstention. Usually, however, two or more options are available, and any vote for one of these alternatives in the set being offered is taken, as an abiding tenet of common sense, to reflect the fact that this alternative is seen by the particular voter as more palatable--in whatever sense--than the alternatives left unchosen.

Such a tenet is so obvious, and so close to the tautological, that it would not strike most observers as either exciting or at all enlightening. Nor can it be claimed that recent decades of serious research into the dynamics of popular voting have in any strict sense called this tenet into question. Nonetheless, this tenet comes freighted with a number of loose corollaries that tend to be taken for granted, and some of these corollaries have in fact been cast in a strange light by intervening research. What is important for our immediate purposes is that the evolution of conceptual frameworks for understanding voting behavior can be most efficiently reviewed if this tenet is kept in the foreground.

While inquiry into voting processes is not new, it is only in the past several decades that such a tenet and its presumed corollaries have been subjected to close examination. That is, inquiry as to the mathematical properties of voting systems has gone on for more than two centuries (see, for example, Condorcet, 1785), but in such deductive work the nature of the vote as a preferred or utility-maximizing alternative is built in as an assumption from the start and is not in question. When ambitious empirical work on voting in large-scale populations began to be undertaken around the turn of this century, the raw data were aggregate voting statistics that permitted no glimpse of the motivations of individuals in forming their decisions. Thus for example, while Siegfried (1913) marvelled at the century-long constancies in geographic patternings of the popular vote in plebiscites and legislative elections in France and attempted to account for them in terms of factors like temperament and land tenure patterns, the basic nature of the vote as a preference expressed among alternatives was taken for granted.

The Columbia School: The Vote Decision as Brand Preference or Social Fact

The earliest large-scale sample survey studies of voting, conducted as basic academic research by Lazarsfeld and his associates at Columbia University, were designed with this commonsense tenet and its numerous intuitive corollaries firmly in mind; but the new information provided by extended individual-level interviews led to a series of empirical surprises. Lazarsfeld had been engaged for some years in scrutinizing the effects of mass media advertising campaigns on consumer preferences. Foundation funding made available for an elaborate study of the 1940 presidential election led him to transfer the media influence model to the arena of democratic politics (Rossi, 1959).

That people would cast votes to express preferences among alternatives was not, of course, at issue. The key research question had to do with the processes whereby the media purveyed details about the alternatives over the course of the presidential campaign, and the voters weighed these details and arrived at their final voting decisions in much the same way as consumers were presumed to digest the competing claims of advertisers in order to arrive at a decision about a brand to purchase. Since the process of decision making was central, Lazarsfeld pioneered a longitudinal panel design, selecting a sample of voters in Erie County, Pennsylvania, to be interviewed starting in May, well before the beginning of the main presidential campaign, then reinterviewing them at monthly intervals through the election in November. A major parallel investment was made in monitoring the output of the mass media in the area in order to be able to match peculiarities in information flow with twists and turns in public evaluations of the competing candidates.

As it turned out, however, the data were hardly as expected, and the elaborate design was found to be a serious misfit for the subject matter at hand. Instead of tracking an electorate through its stages of forging personal decisions as to which candidate to support, the investigators discovered that a remarkable majority of the voters surveyed had already made up their minds about their November vote before the time of the first interview in May, which was in turn several months before either major party had arrived at its presidential nomination.[1] All told, less than 10 percent of the panel being reinterviewed showed change in presidential preferences during the course of the campaign, a number far too small to sustain the kinds of sophisticated analyses that Lazarsfeld had envisioned.

Of course nothing in these results by themselves would serve in any strict sense to disconfirm the postulate that individuals select among alternatives and use their votes to express preferences among them. However, it rapidly became clear that this commonsense tenet did in fact carry with it a number of almost irresistible companion assumptions that the data were calling into question in the most direct way. Instead of progressively weighing the competing candidacies as the campaign highlighted their differences, much of the public had decided in effect for whom to vote months in advance of the alternative candidates' actually being selected. Moreover, it turned out that the small minority who either showed change during the campaign or at least refrained from foreclosing on a decision until very late in the campaign, hence fitting the commonsense phasing assumptions with which the investigators had begun the study, more dramatically violated other cognate assumptions. Lacking a clear preference among the alternatives until deep into the campaign, this small group of voters was most in need of the kinds of clarifying information the campaign would provide. It was thus ironic that they were the least likely of any in the sample to expose themselves to channels of political information during the campaign. The early deciders were much more receptive to campaign information, although it appeared that instead of sifting through a representative sample of partisan claims and counterclaims they paid attention selectively to information that would chiefly serve to reinforce the decision they had already made (Lazarsfeld et al., 1944).

---

[1] Because of informal traditions against a third term, the renomination of incumbent Franklin D. Roosevelt was a good deal less than clear early in 1940, and Wendell Willkie, the surprise nominee as his opponent late in the summer, was almost unknown to the public in the spring, much less seen as a viable presidential candidate.

Thus, while the core tenet that voters evaluate alternatives to find the one best satisfying their personal preferences could be stretched here and foreshortened there to fit the general cast of the Erie County data, it was reasonable to question whether there was anything left to the tenet but the tautology that voters have chosen to do what they wished to do. Certainly the main intellectual meat in the original commonsense assumptions, the notion of the voter as a rational, independent seeker of information and sifter of alternatives, had received an enormous setback: Not only did the Erie County data show few signs of such a syndrome, but also, where obvious predictions could be made from these conventional corollary assumptions, the data often showed significant differences in exactly opposite directions.

There were other surprises in the data as well. Although the investigators had conceived of their voters as individuals digesting information about alternatives gleaned mainly from the mass media, they became impressed in the course of their analyses at the apparent strength of interpersonal influence in primary groups producing partisan homogeneity and, writ larger, the explanatory power of group differences, such as those between social classes, religious creeds, or urban versus rural backgrounds, in accounting for variation in vote decisions. They concluded that voting was less a matter of individual ratiocination than a social fact, and the later work of the Lazarsfeld group on voting in the 1940s was dedicated to the pursuit and elucidation of these themes (Berelson et al., 1954).

The Michigan School:
Candidates, Issues, and Partisanship

The 1950s saw the development of a series of biennial studies of national elections by Campbell and his associates at the University of Michigan. The emphasis in these studies was somewhat more social psychological than had been true of the Columbia work, with a primary focus on the structure of attitudes toward political objects that determined the voters' final decision at the polls. In particular, separate measurements were taken of the voters' orientations toward the major parties, the competing candidates, and the policy debates that seemed most crucial for understanding the election. Much attention was paid to the comparative properties of these orientations and the way in which their interplay contributed to vote decisions, both individually and aggregatively (Campbell et al., 1954).

Although there was some displacement of conceptual frameworks and hence of emphases between the Columbia and the Michigan work, the picture of electoral reality emerging from the Michigan studies of the 1950s looked a good deal more like the world discovered by the Columbia team to their surprise in the 1940s than it did to the cluster of assumptions about popular voting that had held sway before any significant empirical spadework with the sample survey had been accomplished.

This similarity was nowhere more evident than in the contrasts between public orientations toward issues and those toward the major parties, which the more extensive Michigan measurements underscored. The conventional understandings of voting that had originally led the Lazarsfeld group astray through the prototype of the rational independent voter presumed that parties were distinctly secondary to preferences on policy issues. After all, one did not support a political party solely to support it. The rational voter supported it in a particular situation merely as a means toward some more ultimate goals embodied in a program of policy advocacies that were seen as preferable to those being urged by competing parties. Thus the alternatives between which the voter was presumably expressing preferences were policy alternatives. Issue positions were primary, according to the commonsense notion, and preferences for parties per se were assumed to be distinctly secondary and derivative.

The Michigan work of the 1950s called this expected pattern even more sharply into question than the Columbia discoveries had. Much of the public was poorly informed on even the major policy debates in the campaigns, and many voters seemed to have opinions on these issues that were very faintly crystallized at best and hence highly volatile. Moreover, there was a good deal of confusion as to which policy positions were associated with which of the two major parties. Typically, in the aggregate there was a mild association between positions and parties that would be recognizable to the sophisticated observer. However, these trends were usually faint, with large absolute numbers of voters making contradictory associations, not to mention substantial fractions of respondents who confessed immediately that they did not know which party was associated with which position on particular issues (Campbell et al., 1960).

What the vast majority of citizens did know was that they were either loyal Democrats or loyal Republicans, and this sense of emotional identification with one or the other of the two parties appeared to have a primacy and an autonomy relative to any concerns about policy programs that flew directly in the face of long-standing assumptions. Little wonder then that in May 1940, when the Lazarsfeld group began its precampaign measurements, an unexpected majority of citizens had already made their vote decisions. Granted the parties had not yet constructed their competing policy platforms to lay before the public, nor had they even begun to designate their individual nominees to be standard-bearers in the campaign. What was predictable as early as May was that a Democratic slate of nominees and platform planks would be run against a Republican slate of nominees and platform planks, and these alternatives were already enough to convince a very large fraction of the electorate as to its choice.

The Columbia group had reluctantly concluded that their panel study had been too late to catch most of the interelection change in partisanship, and they opined that longitudinal studies would have to be conducted earlier in the period between major national elections to permit such change to be studied. In the subsequent three decades the Michigan group conducted two national panel surveys, each one bracketing a full quadrennium, from before one presidential election until after the next one. Although conducted in rather different political climates, the two studies concur in demonstrating that there is no greater change in underlying partisanship in the interelection period than the paltry amount Columbia group found in the six months of the 1940 campaign. Per unit of time, the rate of change seems remarkably constant and, by any expectations, very slight indeed.

If fundamental partisanship is so firmly rooted for so many and shows negligible change over even extended periods of time, how can we explain swings in the popular vote for major offices from one election to the next, occasionally moving the vote division by as much as 10 or 12 percentage points? The answer was obvious in the Michigan surveys: While most party identifiers voted loyally most of the time and were especially sure to do so if they had no particular information about a given contest, nonetheless they could be induced under circumstances of special stress or special attraction to the candidate of the opposing party to engage in short-term defection. The fact of defection was recognized as such by the individual, and the event of defection did not seem to operate as a halfway house to an actual change in partisanship: Once the peculiar circumstances that had originally induced the defection, such as a special pairing of candidates, had disappeared, the voter was very nearly as likely to vote loyally in ensuing elections as fellow partisans who had not defected. Thus, short-term vote swings could be localized as largely a phenomenon of defection, beyond the contribution to voting change from the small fraction of the electorate that was truly independent.

Long-Term Political Change and Partisan Realignment

In the same period of the 1950s when the party-based Michigan model was being explored, scholars working with long-term aggregate voting trends began to appreciate the degree to which these time series for fixed jurisdictions behaved much less like the random walk of the stock market, for example, and much more like the pattern called a quasi-stationary equilibrium. A time path of this type shows a pattern of short-term fluctuation that appears to be essentially random, although the oscillation seems to be taking place around a mean or equilibrium value that is constant in the intermediate term. Now and again, however, at relatively long intervals, it seems as though the equilibrium level is brusquely displaced and a new series of fluctuations begins, looking much like those of the preceding period save that they are now centered on a new equilibrium value. V. O. Key (1955) labeled these points of displacement of the equilibrium in time series of the vote division as "critical" or "realigning" elections.

Although the survey-based Michigan model and the recognition that macrocosmic voting data performed as a quasi-stationary equilibrium over time were quite independently generated, they occurred at about the same time and the fit between the contrasting micro and macro observations was elegant. The random short-term fluctuations in aggregate vote series were the ebbs and floods of short-term defections; and the equilibrium value around which these fluctuations occurred marked the basic division of party loyalties in the jurisdiction. Although no realigning election had occurred within the span of time covered by individual-level sample surveys, the implication was strong that the rare displacements of equilibrium levels betokened a set of forces to defection that became so strong as to break partisanship, rather than bend it temporarily, and thus to produce large-scale conversion.[2] Thus a micro-level model developed from modern survey data helped historians to understand the character of fine-grained processes underlying major political changes in the 19th century.

The general model was also used as a solution to the long-term riddle as to why the party occupying the White House had for 100 years almost invariably lost seats in Congress in the ensuing off-year election. Politicians had long been aware of the regularity, but the only explanation for it--that the sitting president was bound to alienate some support in the two years after election--was exceedingly lame in view of the fact that running as an incumbent two years later he could expect to have not only his original support back, but usually more as well. Campbell (1960) showed that the phenomenon was a compositional matter, traceable to differences in participation between high-turnout presidential elections and lower-turnout off-year elections. Persons participating in presidential elections but not in off-year elections tend to be less stably partisan than those voting faithfully in both types of elections. They register disproportionately in the surges of defections important in the election of presidents. In the off-year, however, they disappear from the active electorate, leaving the field to more stable partisans whose vote division will under almost all circumstances lie closer to the equilibrium value of the national vote division than did the vote at the preceding presidential election. The Campbell mechanism, once understood, not only served to explain the regularity itself but also was able to predict rare exceptions to the regularity two years in advance of the off-year election.

---

[2] More recent research has introduced some technical modifications in this final stage of the account, although the spirit of the account remains valid.

Rational Choice Models of Voting

More recent years have seen a considerable resurrection of interest in more "rational" or issue-oriented models of voting behavior, which in effect seek to lead back to a view of voting more firmly centered than either the Columbia or the Michigan renditions in the original commonsense tenet that citizens do, after all, cast votes that express preferences among available alternatives. It is not easy to describe participants in this reaction as a single school, because the rubric *rational* brings together strange bedfellows whose interests and procedures are contradistinct. At one extreme are scholars with little familiarity or interest in the details of real-world voting systems, who are mainly eager to press forward with mathematico-deductive work on the logical implications of abstract voting systems under conditions manipulated by mental experiments. Scholars at this extreme adopt the "rational-choice" label because all of their formal modeling begins with the assumption that their hypothetical voting systems are populated by citizens who vote for alternatives closest to their own ideal. At the other extreme are scholars without any apparent skill in formal modeling or interest in deductive approaches who do have an inductive interest in demonstrating that the view of the voter developed in the 1940s and 1950s as remarkably ill-informed and more responsive to group identification in voting choice than to policy concerns was substantially exaggerated. For those of this persuasion, the voter is typically much more rational--i.e., issue-oriented--than earlier work would suggest.

These are, of course, merely extremes described to help convey the heterogeneity of the rational choice umbrella, and hence the difficulty of doing justice to these trends in a single account. An increasing number of scholars in the field combine both deductive formal modeling skills and inductive fascination with problems in the interpretation of real-world systems and have a specific interest in drawing the power of deduction and the insights of induction more closely together in voting research. We describe some of this convergence at the conclusion of this section. First, however, let us begin with the deductive and the inductive streams as distinct.

The Growth of Deductive Modeling

The event that touched off a torrent of deductive work in the 1960s and 1970s was the publication of *An Economic Theory of Democracy* in 1957 by Anthony Downs. Downs erected a structure of assumptions about voters and candidates as actors and proceeded to deduce from it several interesting features of system performance that could be expected as a result. Borrowing from work on location theory developed in an economic market setting by Smithies (1941) and Hotelling (1929), Downs used as a central vehicle a unidimensional (ideological) continuum along which voters were arrayed and on which candidates might position themselves. The voters were rational in the sense that each could be counted on to vote for whichever candidate was positioned most closely to him or her. Candidate behavior was governed purely by a motivation to maximize votes, rather than to promote political principles. Under a set of plausible conditions defined by such things as the shape of the distribution of voters on the continuum and the circumstances under which nonvoting would occur, Downs showed that candidates would, in "me-too" fashion, take up positions very close to one another at the center of the ideological continuum.

Downs explicitly denied any interest in the real-world relevance of his model: It was to be taken as a normative model aimed at saying what would happen in a system if all component assumptions held, without any pretense that they would in fact hold jointly in any real system. However, the pressures toward centrism were familiar enough within two-party politics to be thoroughly evocative. Moreover, while the scheme was obviously a considerable simplification on reality, as any abstraction is bound to be, it seemed bent toward reality in a number of ways. Thus, for example, while the voter as well as the candidate was designed as "political man" in the image of the "economic man" of much macroeconomic theory, there was an excellent discussion of information costs, and the information required of the voter was kept light out of apparent respect for empirical findings about information limitations in the electorate.

In the intervening two decades a wide variety of work has flourished from this original seed. Some of it has been purely deductive. The basic Downs model has been more explicitly mathematicized, and its generalization from the original unidimensional version to n-dimensional spaces has been completed (Davis et al., 1970). Other work shows clear signs of yearnings toward verisimilitude. Much experimentation has gone on with respect to the continued mathematical tractability of the model under partial or complete relaxation of some of the less plausible assumptions. Implications of the model have also been examined for variant real-world contexts, such as optimal strategies for candidates who must first compete in party primaries to win nomination, then, without too much repositioning being permissible, must go on to compete in the general election.

Apart from its impressive deductive contributions as a normative model useful as a baseline for observing deviations in real-world voting systems, the rational choice approach has both strengths and weaknesses in addressing reality. These are perhaps clearest in its treatment of the phenomenon of abstention or nonvoting. On one hand, although the problem has been wrestled with repeatedly, it seems impossible to explain from the pure utility-maximization assumptions of rational choice theory why, as voting systems attain much size at all (millions of eligible voters), it is "rational" for any given voter to bother voting at all. Interested scholars have either chewed on the problem and given up or have been obliged to solve it in ad hoc fashion by retreating to more social psychological terms, prominent in the Columbia and Michigan work but quite alien to utility-maximization conceptions, such as human expectations, human values (e.g., in this context, the sense of civic duty), and patterns of habit formation. On the other hand, once this initial obstacle is ignored, as reality requires it to be, then developments on the Downs model have made a major contribution to the conceptualization and the operational measurement of nonvoting. It turns out that model deductions are highly sensitive not only to the possibility of nonvoting as an option, but even more especially to what kind of nonvoting the abstention represents. It matters greatly whether it is due to indifference or to alienation and furthermore how elastic the turnout decision is as alienation increases. The model is crystal clear as to how such differences may be measured. Here, as at a number of other points, concerns and discriminations that have come to be highlighted by progress in the deductive stream have had a visible impact on measurement practices in voting studies on the empirical side of the watershed.

Important as the Downs model has been, it is not the only source of deductive stimulation to voting research in recent years. In a utility-maximization mode, problems of the transitivity of preferences at the individual level, and more especially the potential for cyclical majorities (see the Arrow theorem: Arrow, 1958) at the collective level, have attracted a good deal of attention. The real-world spin-offs range from efforts to estimate institutional conditions affecting the actuarial probabilities of cyclical majorities to practical suggestions for revisions of traditional conceptions as to the form in which votes are solicited (preference voting, approval voting, and the like) in order to clarify true voter preferences under circumstances in which more than two alternatives are offered (see the summary discussion in Gardner, 1980).

The Inductive View: The Issue-Oriented Voter

The Downs model was extremely attractive to many scholars energized less by an interest in formal modeling than by distress at the empirical image of the voter emerging from the Columbia and Michigan work, which tended to downplay the importance of ideology and policy issue concerns in favor of group identifications in general and partisan identifications in particular. This was so because the spaces in models of the Downsian type were typically conceptualized as being defined by dimensions of ideological differentiation or more specific policy cleavage. Hence to think in terms of such models was to restore policy preferences to their rightful place in the voting process.

This association was, strictly speaking, something of a happenstance. Not only did the original Downsian impetus disclaim any pretense to verisimilitude, but also there is nothing in modeling of the Downsian type that requires the spaces to be defined in policy terms. As several scholars in the Downsian tradition have pointed out, the machinery would proceed in exactly the same way if some or all of the relevant dimensions along which voters and candidates positioned themselves were defined in terms of relative partisanship or the personal charisma of candidates.

A much more pointed catalyst came from a posthumous work of Key (1966), which argued that voters were about as responsive to issue concerns as the system (including the mass media and the degree to which candidates were willing to differentiate themselves in issue terms) permitted them to be. The work gained considerable force because of accompanying empirical analyses, which implicitly represented a shift in the research agenda. Instead of attempting to account for variation in the total vote, these analyses examined change in the vote at the margin from election to election. This much more restricted variation set aside by definition anything that might be explained by an inertial term like partisan loyalty. With the party term ruled out, Key was able to show that a rather impressive portion of the residual (change) variation was correlated with voter concerns about prominent issues surrounding the election. The shift in research agenda was, of course, entirely warranted in view of the conceptual importance of voting change in determining the ebb and flow in party governance in the country.

Time series of aggregate vote statistics by definition measure exactly the same kind of change at the margin that occupied Key and in many ways offers a firmer base for analysis because of the potential length of such series. Several scholars, including Kramer (1971), have shown that a number of highly coherent patterns can be found in aggregate data linking fluctuations in economic welfare with shifts in the aggregate vote, much as the Key argument predicted. On one hand it seems generally true that when voters are regarded individually, one is most apt to be impressed with things like low levels of information and highly idiosyncratic constructions of political reality. On the other hand, it is equally true that when one focuses on change in party fortunes from election to election with highly aggregated data, much of the large admixture of "noise" represented by diametrically opposed perceptions visible at the individual level washes out through mutual compensation, and what is left as residual net change, even though it may rest on the assessments of a very limited fraction of the electorate, does seem to express overall messages that are safely construed in terms of policy or substantive grievance. Thus the aggregate system is a more "rational" channel of communication from the grass roots than work with individual data might suggest.

At the same time, another solace concerning the relation of popular voting to issues has been found in what seems to be some real change in attentiveness to issues in recent decades. Survey data on voting from the turbulent politics of the later 1960s and early 1970s look discernibly different from those of the war-occupied 1940s and the quiescent 1950s on which the first fine-grained portraits of voting behavior were based. Among other trends the impact of party loyalties on vote decisions has declined somewhat, and issue sensitivities have shown some increase (Nie et al., 1976). These shifts at the individual level are only relative and have certainly not reversed things. Moreover, it remains a moot point whether these trends are truly secular and will continue, or whether they represent segments of more nearly cyclical fluctuation. There is beginning to be empirical reason to make the intuitively plausible links between the sharpness of policy differentiation between parties and their candidates and the degree to which voters are influenced by policy concerns in forming their decisions at the polls. In any event, the continued study of voter behavior over a lengthening array of elections, and hence variety of social and political contexts, has become increasingly helpful in understanding the range of colorations that may be detected in popular voting.[3]

Models of the Fuller Voting System

Whereas a great deal of effort has been dedicated to discovering what makes the average voter tick in forming political evaluations, inquiry in this area has steadily broadened in the past decade or two toward a much more inclusive view of the political system. The Downs model itself represented an early stage of this broadening. Realistically, the model was scarcely one representing voter decision making at all. The voter was typified by a few behavioral assumptions; and voters collectively were considered as distributions of opinion, the shape of which was taken as fixed or given in particular applications. The moving parts of the model involved candidates and their maneuvers in competition to adjust optimally to whatever distribution of opinion was being hypothesized at the time. Thus it was more a model of candidate strategies than of voter behavior, although, since the voters formed the crucial environment to which the candidates were obliged to adapt, it was perhaps most aptly described as a model of candidate-voter interaction.

---

[3] Although we do not develop the subject here, it is also true that collection of parallel survey data on voting from an ever-widening list of democracies of varying economic development and political institutions (different party systems, different configurations of laws and procedures surrounding popular elections, and the like) is beginning to provide some initial sense as to which behavior patterns are most generic and which are shaped in one way or another by the developmental and the institutional context.

These broader interactive models, which have been increasingly exploited, tend to hinge on rather different conceptual apparatus than is most natural for questions of determinants of the voting act, although conceptions of the popular vote must obviously inform such models. Focus in these models may vary, although the common citizenry usually figures as one party to the interaction, with some kind of political elite figuring as the other party.

In the Downs vein the most frequently studied political elite in such interactive models is the set of candidates for office, and the practical question at stake is indeed one of how to maximize the likelihood of winning an election. A variety of studies, touched off by Pool et al. (1964), have dealt with the empirical relationship between apparent distributions of voters on particular issues and optimal responses of the candidates, not only in positioning themselves on the issues but also in selecting policy areas that are advantageous for them to emphasize. Lore of this kind, fueled by sample surveys conducted by campaign teams tailoring their work to a specific jurisdiction and its relevant issues, has become almost a staple of any major electoral contest. Meanwhile on the deductive side, study has gone forward in a number of areas ranging from optimizing the allocation of campaign resources (see Kramer, 1966) to questions about the impact of relevant formal models of various tactics, such as obfuscation of position by the candidate so that his or her "position" will be perceived at best as a range of greater or lesser width on the policy continuum.

Another political elite drawn into these interactive models is the subset of candidates who win election and thus become the representatives of their constituencies. Normative theories vary widely as to the exact performance obligations of representatives (Pitkin, 1967), and this spectrum of theories is reflected empirically in the variety of conceptions as to the appropriate role vis-à-vis the constituency that may be found in any real-life legislature (see Wahlke et al., 1962). But however strict or loose it may be, that there is some obligation to the constituency, backed by the threat of defeat at the next election, is hard to question in a democratic society.

Although most deductive work has stopped at the level of candidate strategies, inductive designs have been developed for the examination of the representative-constituency interaction and have been implemented on an increasingly broad scale. National-level versions of such work typically follow the Miller-Stokes design (1963). A sample of legislative districts is drawn and their representatives in Congress are interviewed with regard to their personal positions on central policy issues of the day as well as their perceptions of public sentiment on those issues in their constituencies, their views as to the appropriate role for the legislative representative, and the like. Within each of the selected districts, samples of constituents are independently interviewed with respect to their actual positions on the same issues and their perception as to their representatives' positions on these issues, along with many other kinds of information normally elicited in such a survey.

With such data from constituencies and from their particular representatives, a variety of important estimations can be made. Thus, for example, representatives' guesses as to district sentiment can be compared with direct measures of such sentiment and thereby graded for accuracy, just as the voters' perceptions of their representatives' stands on the issues can be compared with those actually reported by the representatives themselves. In the standard version of the design the representatives' performances in the legislative body, as embodied in such summary output measures as roll call votes on legislation relevant to the measured issues, is also monitored. While such legislative votes are subjected to a myriad of influences and pressures, and the legislators' own opinions on the relevant issues predict their roll call votes only imperfectly, the configuration of information available permits a wide range of important assessments to be made. To the degree that representatives depart from constituency sentiment in their own legislative voting in some policy domain, is it because they do not see their role as requiring any faithful representation of the postures of their constituents or, although wishing to be an instructed delegate, do they perceive opinion in their districts inaccurately? Or again, are they responding not to their constituency as a whole but rather to some subset of it, such as constituents of their own party, those who voted for them, or some particularly vocal and articulate minority?

A still fuller version of the design is focused on a specific election, with all candidates for the legislative seat being interviewed rather than just the winners alone. Such a design allows comparisons in yet another direction: How do those candidates preferred by voters in the district look on relevant issues when compared with their antagonists from whom the majority turned away at the polls? Do they perceive district opinion more accurately than the losers, less accurately, or is there no difference? Most important, are they in point of fact better representatives of district sentiment than the losers, and if so, by how much? The question of "how much" leads to a crucial index as to the policy significance of the fact that the public is permitted to intrude on elite decision making at all, sifting candidates into winners and losers. Purely and simply, it provides a measure of the degree to which the conventionally assumed function of popular elections is being realized.

Finally, the comparative replication of such a design across democracies organized with substantially different political institutions, which has begun to be carried out in the past 15 years, permits investigators to inquire as to the ways in which such institutional variations affect the process of representation. How much does it matter to popular representation, for example, that in some countries party discipline in national legislatures is virtually absolute, whereas in others, such as the United States, it is significantly less stringent? Is political representation of popular sentiment more faithful in multiparty systems, or less? How do systems with proportional representation or at-large national legislative districts compare with those organized by single-member districts with majoritarian or pluralitarian winners?

It is, of course, a long leap from questions aimed at what makes individual voters tick as they form their voting decisions to questions of the magnitude that can be meaningfully posed at the level of these interactive models of the fuller voting system. But it is this kind of conceptual development that brings basic research on voting to the point at which it can address the vital functioning of democratic practice.

ELECTION REFORM AND ELECTORAL RESEARCH

The interface between practical concerns of governments about the conduct of public elections and the progress of basic research on voting extends beyond legislative concerns over the design of laws to govern that process. Thus, to take recent examples: Testimony from voting experts may be required in litigation in the federal courts in which the failure of a particular voting machine to continue to register the votes entered after a certain hour is alleged to have been critical for the larger outcome, or litigation in which the conduct of an election involving the National Labor Relations Board in a given plant is alleged to have been unfair on one side or the other. The major area of the interface nonetheless does involve the broader question of the optimal design of election laws.

Here, as we have suggested, the traffic between the two worlds is distinctly reciprocal. More often than not, agitation for electoral reform arises from the perception of abuses that grows through government or journalistic observation or from shifts in the definition of national values, as with extensions of the franchise to women or to those under 21. However, such agitation is sometimes fueled by what is going on in basic research on voting, and this source for critiques of existing regulations, existent for only a small fraction of the nation's voting history, is likely to increase in the future. Moreover, as the possibility of particular reforms is debated, basic research is consulted for evidence of the likely impact of the change, including not only its efficacy in producing the desired result but also its capacity for unforeseen consequences. Once a particular reform becomes law, researchers are likely to descend on data generated under the new arrangements, after the fashion of the natural experiment, in order to assess its consequences more directly. That other social, political, and economic circumstances may have changed as well as the election laws is a confounding element, as it is with any natural experiment, although the growing lore as to what "normal" temporal variation in the fine infrastructure of voting patterns looks like permits increasingly incisive inferences about the likelihood of confounding intrusions. In short, basic research can stimulate electoral reform and help to assess the value of proposed remedies; and in turn the fact of such reform creates a source of institutional variation that basic research can exploit to raise the level of generality of its own propositions.

It is unlikely that there is any unique optimal structure for election laws in the context of large national states, and it is understandable that concern with electoral reform never dies. It waxes and wanes with changes in what democratic theory or the democratic faith expects of citizens, representatives, and the relationship between them. Remedies, more often panaceas based on intuition than antidotes based on knowledge, are easy enough to prescribe. American electoral reform has been a continuous process because almost every reform has had its unanticipated side effects that, in due time, are themselves in need of remedial action. This is as one might expect, for genuinely novel reforms in social institutions and processes are nonexperimental undertakings in which the test comes in the course of the remedy's administration rather than in some prior demonstration.

Consequently, almost all of the reforms ever instituted in American electoral practices at one or another level of the federal system have been interpreted as both successes and failures, depending on circumstances and expectations that change over time. One can cite various extensions of the franchise; the adequacy of rules designed to protect the secrecy of a vote; the use of long versus short and party-column versus office-column ballots; the introduction of initiative, referendum, and recall elections; the application of majority voting, plurality voting, or proportional voting as decision rules; the institution of the electoral college as a screen between popular vote and actual decision; the institution of permanent registration and other means to ease registration; the employment of primaries and runoff elections; the formation of single-member versus multimember constituencies; revisions of the electoral calendar favoring staggered over unstaggered elections; the control of bribery, fraud, and other abuses in polling and the abolition of poll taxes; the limitations placed on campaign contributions and expenditures; the enforcement of fair campaign practices; and so on.

Political scientists specializing in voting research have been active contributors to the shaping of this aspect of public policy, in view of their generic concern with the impact of political institutions on individual behavior. Thus, for example, they made substantial contributions to the presidential commission that guided the federal election legislation in 1964 and 1965, based on their rapidly increasing understanding of impediments to voter participation in electoral mechanisms.

In the later 1960s ferment arose from other quarters to lower the national voting age from 21 to 18, producing a spate of dire predictions among opponents that such a reform would lead to a major degradation of the electoral response by inviting hordes of uninformed adolescents into the electoral process before their judgment had matured. Without necessarily taking a particular position on the merit of the change itself, scholars of voting could make a number of predictions as to its effects. There was reason to expect that the new infusion taken alone would add to the responsiveness or elasticity of the electorate to very short-term considerations affecting the vote, although whether such responsiveness should be described as a new flexibility or a new instability was a matter for individual values. More certain and possibly more important, any such effect would be limited in the aggregate, in part because the additional cohorts were such a small fraction of the new whole but more especially because prior research had indicated this fraction would be rendered even smaller than might appear because it was predictable that these new cohorts would show the feeblest rate of turnout save conceivably for the very oldest (beyond, say, 80 years of age). Subsequent experience has borne out all of these predictions.

In like manner the political scientist's understanding of the relationship between one's age or political experience and the stability of one's partisan sentiments has provided a useful perspective on the contribution of the voting age reforms to declining partisanship in the electorate at large. Although the young cohort newly admitted to the electorate with a lowering of the voting age were only a minor fraction of the voting public, the particular timing of their admission, de facto in 1972, affected others of the nation's political parameters in a very visible way. Because the change in voting age occurred just as those in the postwar baby boom were reaching their late teens or early twenties, almost one of every six citizens eligible to vote in 1972 was eligible for the first time. In 1952, with the voting age still set at 21 (except in Kentucky, Georgia, and Tennessee) and the baby boom just under way, only some 7 percent of the electorate were newly eligible to vote in the election of that year; in 1972 the proportion reached 15 percent.

Although less than half of the new cohort actually voted in 1972, they were all included in national surveys and polls that were estimating trends in the partisanship of the electorate. In those estimations they made a disproportionate contribution to the national growth of political independence, with well over half calling themselves some variety of nonpartisan. Following the same cohort over time, political scientists have now documented through recent data an upward surge in their partisanship, a surge that is speeding the restoration of partisanship in the total electorate to earlier levels (Miller and Shanks, forthcoming). These analyses of partisanship have added further evidence to the expectation that the effects of the voter age reforms would not overwhelm or destroy the existing political system.

More recently, in response to an unbroken trend of declining turnout on election day, scholars anchored in basic research have been directly addressing what are conceived to be dangerous electoral malfunctions in need of diagnosis and cure. Thus, for example, Wolfinger and Rosenstone (1980) have been able to use the conceptual frameworks and methodological perspectives of basic research to estimate likely increases in turnout that would result from the elimination of burdens placed on citizens for maintaining their own eligibility to vote in subsequent elections.

Similarly, changes in the practical political arena have led to a resurgence of interest among scholars in understanding the impact of campaign finance on election outcomes (see, for example, Jacobson, 1980). Of course, political scientists have long been interested in the role of money in politics. The early interests were heavily centered on questions of corruption and improper influence of wealthy elite sectors on political office holders. With the development of new campaigning technologies that both provide and consume unprecedented amounts of money, interest has moved to more complex problems. In California, for example, candidates for the 80 state Assembly seats spent an average of $353,000 per seat in their 1980 campaigns; two of the candidates spent more than $1 million each--more than was spent by all candidates for Assembly seats in 1958 (Jones, no date). Such levels of expenditure raise profound questions about the extent to which candidacies for office are limited by access to campaign funds. In the meantime, policies providing public funds for campaign costs have proliferated across the nation. Each state's policy has its own variant on themes of record-keeping and public disclosure, limits on contributions, ceilings on expenditures, and formulas for the allocation of funds.

Political scientists are now adapting their research to respond to needs for more specific policy guidance, providing another set of examples of the interaction between basic and applied research. The practical needs for new policies give added impetus to research on candidate recruitment, on the efficacy of different campaign strategies and techniques, on motives for citizen participation in politics (including participation through giving), on the role and effective functioning of the party organization, and on voter response to election campaigns. What are the consequences of new funding patterns for candidates and parties? Is the political party being removed from the election process, and if so, with what effect on the role of party in government? If the contact between candidate and citizen is more direct and less often mediated by considerations of party, what are the consequences of both the individual and the aggregate vote decision? Will citizens' reliance on costly media and mass communication replace the role of group leadership and interpersonal influence in the voter's decision making process? Can public funding provide the leverage to curtail campaign expenditures, equalize access to candidacy, and restore the institutions needed for responsible party government?

Although the phrasing of the public policy questions tends to carry overtones of normative, value-laden goals, the political scientist's answers are structured by the empirically based, theoretically organized research that is the basis for diagnosis and prescription. Although assessments of this kind may be seen as practical or applied research, the borderline between such activities and basic research in the area is faint and probably useless to insist on because the interpenetration of basic research concepts and empirical baselines with practical concerns is virtually complete, and the substantive returns of such investigations to basic understanding of the "political man" are considerable.

THE GREAT EXPANSION OF COMMERCIAL POLLING AND VOTING RESEARCH

To the degree that voting research has increasingly replaced guesswork and intuition about possible variations in structure of electoral law with a corpus of fact and understanding, the social utility of the research contribution is not very controversial. The role of commercial polling in the electoral process has, however, been a subject of controversy almost from the outset, particularly because it involves extensive dissemination of findings in the mass media, continuing projection of election winners, and the like. We first review the history of the considerable interpenetration between basic research on voting and trends in commercial polling, then move forward to assess some of the more current controversies over the role of the polls.

One major area of the interpenetration is a version of technology transfer, the flow of methodological developments from one of these worlds to the other. As we have already noted, the tool of the sample survey itself moved into high public prominence and received its first broad-scale academic attention through the early commercial polls. Since the 1940s, however, most improvements in methodological sophistication necessary for upgrading quality in the work of the polls have come from basic research. Especially where politics and public opinion are concerned, the work of the commercial polling organizations is permeated by the theories, methods, and analytic techniques that have been developed in basic research.

The most dramatic instance of such flow, as we have seen, occurred in the wake of the 1948 presidential election, through the elaborate inquiry conducted under the sponsorship of the Social Science Research Council (Mosteller et al., 1949). A heated debate had been under way in the 1940s between the pollsters and researchers (including statisticians) in university centers of survey research over the necessity of strict probability sampling procedures. While the report identified a large range of problems in the procedures being used by commercial polls, including the termination of monitoring well in advance of the 1948 election, it gave special emphasis to problems of sampling. It is ironic to note, with the benefit of three decades of further monitoring of elections, that public assessments of competing presidential candidates can in fact swing quite markedly in the very late stages of a campaign (with the 1980 case appearing to constitute one of the more dramatic examples), so that the 1948 problem may well have been merely a matter of having ceased taking the public pulse too soon. Be that as it may, the 1948 experience convinced people in both worlds of the desirability of using some version of multistage area probability samples in lieu of the less expensive but more uncertain modes of sampling that rely on interviewers selecting respondents to match preestablished quotas. While the shift in procedure in commercial polling agencies was very far from complete, either across agencies or across all studies conducted by given agencies, the wake of the 1948 election did result in a major overhaul of polling methods.

The election of 1980 provides another instance of an apparent mismatch between preelection poll results and the actual election outcome. It is still not clear whether the frustrations of the pollsters who missed predicting the Reagan electoral landslide will be addressed by another large-scale inquiry into the methods and techniques of the polling industry. On one hand, some polls were relatively accurate. Others attributed their forecasting problems to major last-minute changes in voter intentions.

On the other hand, at least one of the major polls, directed by Warren Mitofsky for CBS and the New York Times, has produced data suggesting that the accuracy of some polls was only superficial and that deficiencies in polling methodology actually obscured the magnitude of the real last-minute change in voting preferences in 1980. This argument implies that many commercial pollsters did not heed the general lesson of 1948 and should have paid attention to niceties of method more often associated with basic research. It seems more than ironic that the Mitofsky argument is addressed to possible failings in sampling procedures associated with telephone interviewing, a burgeoning technology that we review below.

Some argue that the precision of estimates needed for most public opinion polling does not warrant the expense of strict probability procedures in sampling, even if it were logistically possible to implement them in the constricted time frames for data collection available to the polls, which it usually is not. When this lowered precision is a necessity, it should not be forgotten as it often is when poll results are disseminated to the public (see below). And even a forgiving viewpoint about polling procedures should recognize that the need for precision of estimates for scientific work often justifies the greater expense of more elaborate sampling.

Although there are no comparably dramatic incidents to mark the impact of basic research on the commercial polls where question wording and interview construction are concerned, it seems nevertheless true that concern with problems of validation has been responsible for a number of contributions of the scientific community to the polls.

With reasonably elaborate theories giving impetus to the research, investigators of electoral behavior have given detailed attention to variations in meaning that can be attached to item responses. The theme of multiple methods and multiple techniques for triangulating over potential variations in meaning has become an important one in voting research as in other fields of social scientific endeavor in recent years.

With the staggering increase in the cost of research based on personal interviews of probability samples of the electorate, there is currently a great deal of ferment, both methodological and substantive, surrounding the use of telephone interviews. The development of efficient techniques for random-digit dialing promises to bypass much of the expense of area probability sampling. There are unusual economies that follow from the use of the telephone interview and it seems likely that despite problems of bias stemming from telephone ownership, other sampling problems are at least as easily overcome through the use of the telephone as through elaborate field sampling procedures.

This developing technology, along with computer-assisted interviewing in which responses are directly entered into the computer, has no distinct origin in either the world of commercial polling or that of academic research. Nonetheless, a conversion toward the new technology has been more rapid in the commercial world, where sheer cost and speed considerations have greater priority. Much more careful examination of the relative reliability and validity of the technology has taken place in academic research and is continuing, although, with one class of exceptions, the current evidence seems to suggest that there is little measurable decline in apparent reliability when telephone interviewing is used.

The class of exceptions involves measurement techniques dependent on visual aids to maximize the clarity of the information elicited from respondents. If anything, there has been a drift toward increased use of such mechanisms in recent years because of evidence that they increase reliability of response. However, they cannot be included in methodological comparisons between personal and telephone interviewing, because the telephone medium precludes their use. This is a problem that in the long run may be solved by televisual communication tied to telephoning. In the interim, however, there may be a dilemma similar to that faced by researchers who must choose between probability sampling and quota sampling.

The personal interview conducted with the heavy use of visual aids for the respondent may be justified only by the need for the utmost in reliability and validity of the data.

As we move from method to substance, political substance in particular, it is worth recognizing that while there is an unquestionable affinity between basic scientific research on voting and the applications carried out by the commercial polls either to serve candidates in private or to attract interest through mass media dissemination, it is important not to confuse the two. Their functions are quite different, and it would be harmful to pass judgment on basic research by inference from the polls.

The polls are under enormous pressure to complete their fieldwork rapidly, even at considerable sacrifice of sampling and data quality; otherwise the results lose their timeliness, a matter of primary concern whether the purchaser be a private campaign strategist or a news medium. The items that they include in questionnaires are much less likely to be constructed with any carefully specified theory-derived variables in mind and for obvious reasons are shaped and restricted by the immediate interests of their clients--of candidates who want to know how they are doing and of newspeople who define what is newsworthy and hence appropriate as targets of the polls.

Similarly, the actual use of the data generated differs markedly. In the nature of things, polls are superficially analyzed--hastily interpreted, often by workers whose familiarity with and sophistication about such data are relatively limited--then largely forgotten, their function achieved. Basic research data on voting are subjected to much more lengthy and sophisticated analyses, and their interpretation can be readily subjected to hot dispute by all the self-corrective mechanisms for which scientific inquiry is noted.

Social scientists, like their colleagues in the natural sciences, are under obligation to make their theories and procedures public, and they are expected to make their raw data available to other scholars, at least once findings are reported and often before. The commercial and media pollsters are not so obligated. In the case of polls conducted for candidates or campaign organizations, keeping the poll results secret is considered not only necessary but also ethical. If such proprietary information is leaked, it is often with highly questionable interpretations that are part of the tactical warfare and an obvious pathology of the information process. The results of some candidate polls have been more completely reported after an election, but we are not aware of the actual data having ever been made available for secondary analysis. The media polls report results, tending to let the chips fall where they may, and some of them (though by no means all) deposit the data in archives for scholarly use; but often the necessary sampling and technical documentation is either absent or inadequate.

Thus, while it is certainly not the main purpose of basic research in voting and public opinion, one clear function of such work vis-à-vis commercial polling is to help keep it honest by providing data that on balance are of higher quality and have been more thoughtfully analyzed. Of course, this function is only long-run; short-term correction of misleading poll conclusions from basic research is, save for rare and fortuitous occasions, largely out of reach. Under such circumstances, the damage may already be done before basic research becomes involved.

INTRUSION OF THE POLLS AND MEDIA

All of which brings us to the key area of controversy surrounding the polls, particularly in that part of their political operations that involves the rapid dissemination of results to the public.* What has persistently stirred comment and complaints from the earliest days of polling is that dissemination of public opinion and voter intention information may represent an artificial and unfair intrusion on the modern electoral process.

Some of the suspicions and allegations are small-bore and highly specific, although nonetheless important for their narrow targets. Most noteworthy here is the concern that diffusion of poll results showing candidate standings well in advance of elections is bound to exert unfair influence on the assessment processes of the observing voter, and hence can either foreordain or at least affect the subsequent outcome. Of course pundits have ventured guesses and estimates about which candidates are out in front in American election campaigns from the beginning of things, but what is new is the patina of authority conferred by claims of being scientific as well as the crude convergence of results across a multiplicity of polls. Other allegations of unfair intrusion are considerably more broad-gauge, such as claims that the television presentations of campaigns and other facets of the political process are progressively redefining, in pervasive and pernicious ways, how Americans conceive of politics or how they proceed to evaluate candidates for office.

---

*There is little parallel social concern about their more proprietary operations, since if a campaign strategist hires work of poor quality or grossly misreads the data collected, one can conclude that the candidate is merely getting what he or she deserves. It may, however, remain of concern that polling is expensive, even in its cut-rate forms, and its growing indispensability in campaigning is a significant part of those spiralling costs that mechanically exclude some aspirants for office and at least disadvantage others.

It goes without saying that it is easier to assess putative effects that are reasonably clearly specified, such as those surrounding specific election outcomes, than it is to address the possibility of displacements that are by contention more amorphous and atmospheric, although serious scholars do attempt work at both levels and in between. Even so, the most focused and specific allegations are not any more subject to definitive verdicts than is true of many other subject matters, primarily because of the infeasibility of strict experimental controls. At the same time, some further difficulties arise because contrary directions of suspected influence are often intuitively plausible.

Some of the difficulties are epitomized by the sequence of events surrounding the 1948 election. Before election day there was a storm of indignation at the polls' prediction of a sure Dewey victory, on grounds of the self-fulfilling prophecy: The implicit model was that Truman supporters would be discouraged and would not bother to vote, whereas waverers would wish to jump on the bandwagon and have the thrill of supporting a winner, with both trends simply adding to the expected Dewey margin. The Truman victory naturally stilled much of this comment, which turned instead to complaints about the inaccuracy of the polls. However, logically it was possible to argue, as some diehard commentators did, that the implicit model was correct and Truman would have won by a landslide without the poll intrusion; or that the implicit model was wrong and what the polls had generated was stay-at-home complacency among Dewey voters and a heightened and unnatural mobilization of Truman supporters. Actually, the sparse sources of national survey data tracking the 1948 campaign right up to the election (see Campbell and Kahn, 1952) give little sign of either kind of intrusion, although they can scarcely claim to be definitive in this regard. Nonetheless, with all sorts of post hoc interpretations available--including the one that voters in 1948, unlike those of today, still distrusted the accuracy of polls and hence disregarded them--it is obvious why assessment of this kind must ultimately rest on relative plausibilities rather than proof positive.

A cognate problem involves the concern that election night projections by the television networks, based on data from the East and Midwest, contaminate votes cast on the West Coast where polls are still open. In 1964 the ABC network, stung by allegations of such contamination, commissioned a small survey by independent academic researchers to assess such effects. The study was too limited in size to be very incisive, but it was nonetheless instructive. It is often easy to forget what a small fraction of even the West Coast population in fact votes after 5 PM, or what a minuscule portion of the total national electorate these voters represent. The national projections are made early only when a landslide is in the making elsewhere in the country, and of course these are the cases in which the West Coast contribution is least likely to affect the outcome. But even when there is landslide news and projections are made so early that a significant population is still left to vote on the West Coast, as was the case in 1964, there are further limitations on likely effects. It turned out that in the survey, among those voting late enough to have been able to hear the race "called" before they themselves voted (many of whom were on the way home from work), the fraction that had become aware of the news before voting was smaller than one might guess. It was only this vanishing fraction of the full electorate that was at risk of contamination from the projections, and hence it was within this group that serious work as to detection of influence could begin. However, these cases finally at risk represented such a tiny handful in the sample that coherent analysis was out of the question.

The same general problem arose in still more pointed form when, on election night 1980, President Carter conceded defeat in a nationally broadcast statement before the West Coast polls had closed. This was distinctly more galvanizing news than network projections, and it probably was diffused much more rapidly. It also would have struck its audience as more authoritative. There were severe complaints from many West Coast jurisdictions that vote turnout fell off abnormally after Carter's statement was broadcast, thus intruding on the outcome of local races, if not affecting the national decision. Most crucially, 1980 was sharply different from 1964. Both the concession speech and the network projections of the Reagan electoral landslide were unexpected news on election day 1980, while election day projections of the Johnson landslide in 1964 simply confirmed widely held expectations created well before election day.

On the eve of the 1980 election most polls described the election as "too close to call," and at least one major East Coast daily persisted with a prediction of a massive Carter victory. Late on election day, as the networks' state-by-state estimates of the popular vote cumulated to a national landslide for Mr. Reagan in the electoral college, their projections of that outcome, capped by Carter's early concession speech, were really news. As such they had the potential to alter the voting intentions of those citizens, including a small but visible percentage, in the East and Midwest as well as in the Mountain and Pacific time zones, who had not yet carried out their intention to vote.

Even though the 1980 national election study had some of the same limitations of sample size that had plagued earlier studies on this topic, the numerical limits were not as severe as in the past, and the longitudinal study, with its copious before and after measurements, facilitated a more incisive search for effects. This time, the more nearly ideal conditions yielded research findings that seem quite clear-cut. Using the results of prior basic research to estimate the probability that each citizen would turn out to vote if unmolested by dramatic news of a concession, the impact of that news on actual behavior could be assessed. The results were unmistakable: Those who heard either the concession speech or a network projection of the electoral landslide before they had actually voted but while the polls were still open turned out at rates well below those predicted from preelection study data. News that the election had been decided clearly depressed turnout among the minor fraction of would-be voters who had not yet voted. And given some indication that the voting intentions of citizens who preferred Carter were more often changed by the news, it seems likely that the outcomes of closely contested local races, especially in the West, may indeed have been influenced.

As with the intrusion of poll results in John Anderson's access to public campaign funds, the intrusion of the polls and countering concessions and projections on the actual election process itself constitute a new problem for public policy, a problem that is normative in nature and that can therefore be illuminated and informed by more basic research, but not a problem to be resolved by that research.

On a different canvas, involving elite decision makers rather than the common voter, there is some reason to imagine that misreadings of poll data occur from time to time at high levels, either because of casual interpretations passed on by pollsters or because question wordings have ambiguities that lay readers are unlikely to notice. These are distressing possibilities in view of the power of those absorbing the information in correct or incorrect ways. From one point of view, these difficulties may be taken as nothing that is unique to poll data: After all, being a political leader or a scientist or an intellectual is no proof against forming interpretations of information that subsequent or more thoroughgoing information may reveal to have been patently ridiculous. From another point of view, it is possible that the special frailties and vagaries of information on public opinion are not well understood, especially by those whose contact with them is secondhand. In an earlier era, when poll data were first being introduced, there was a considerable insulation of skepticism, which largely rested on ignorance of things like sampling theory. The pendulum may now have swung the other way and trust may be too great, resting on dutiful acceptance that even small samples can provide reliable information about the whole, but failing to understand the susceptibility of poll data to biases and ambiguities beyond the error margins associated merely with sample size.

THE SOCIAL UTILITY OF RESEARCH ON VOTING

It is not easy to dress a balance sheet that summarizes the social utility of voting research, because the conclusions to be drawn are many and varied. Along the way, as with our discussion of the interplay among concerns over electoral reform, it is hard to resist the impression that information generated from voting studies has at least some constructive social use, although the amount of practical payoff in itself might not convince everyone that the effort was worthwhile.

In point of fact, in many of its facets the advancing understanding of voting generates information that, like many other results of inquiry, is intrinsically neutral--if by that word we mean that it can be put to uses most would see as desirable, but that it can equally well be abused. This fact is nowhere more evident than in the employment of such understanding for competitive advantage in electoral contests. Precisely how potent deductions from professional understanding about voting may be in winning votes is itself undetermined: Certainly much of what voting researchers have learned has underscored the importance of a variety of "unmanipulables," or elements of voter behavior that stand as major obstacles to the candidate eager to move into the scene and in the short term woo large numbers of votes away from the competition. On the other hand, if the art of winning votes has indeed edged in the direction of a science, the relevant research information is available to all, and the monster can use it as well as the saint. Moreover, if both sides to an electoral contest use such information equally, then the matter is a standoff and outcomes devolve once again, perhaps appropriately, to other considerations.

Active abuses of these understandings often lie very close to the surface. If, for example, it becomes clear from basic research that the common voter is less informed on political matters than an earlier period assumed, it is an open invitation to the unscrupulous politician to cut still sharper corners with the truth and to engage more baldly in cynical manipulation. If it becomes clear that the common voter is less attentive to what is going on politically than was hitherto thought, it is an open invitation to politicians to broaden the area in which they feel the latitude to act politically without worry about reprisals from constituents at the polls.

It seems clear at the same time that more exacting revelation of public opinion and the infrastructure of voter preferences from election to election brings politicians into a closer contact with grass-roots sentiment than they have had since people overflowed town meetings, and from the point of view of standard democratic theory this closer contact has to be regarded as a positive development.

It is all too easy to be cynical about the use by candidates of public opinion data to find out where issue sensitivities lie in the constituency in order to cater to them in hopes of winning office. After all, positions may be taken that may be utterly insincere, and the half-life of campaign promises is notoriously short. But suppose we set ourselves back 50 years, before such public opinion information was commonly available. Were politicians then any less eager to cater to public opinion in order to win votes? It seems unlikely. And if motivations were roughly the same, what was the public opinion that was catered to, absent any systematic information from the grass roots? Undoubtedly what stood as constituency opinion was a strange amalgam, heavily weighted by testimony from friends, the posture of the local newspaper editor, and what could be gleaned at the local courthouse and party headquarters. If the candidate cares to conform to public opinion at all, it may as well be to something more representative.

A parallel argument can be drawn for the lengthier and more important periods outside the campaign itself, after a candidate has won and the representation process has begun. Representatives invariably claim that they have an electoral mandate to do what they are in fact doing. But if pressed, they will also admit that their mandate is encumbered by uncertainty with regard to many of the issues on which they are called to make decisions and policies after an election. While their campaigning, visits back home, and communications from constituents give them a sense of what at least some people feel about some issue in a general way, representatives in earlier times rarely had means to assess constituency opinion in any broader way; and without poll information the same uncertainty abides today.

In part, of course, all of this occurs because constituents do not pay close attention to the flux of policy debates or do not have the necessary knowledge to instruct their representatives; in part, such mandate uncertainty also stems from the fact that, in a modern society like the United States, electoral districts are large and populous as well as heterogeneous and composed of often conflicting interests speaking in many voices and making demands on representatives that are difficult to reconcile. Some representatives solve the problem of mandate uncertainty by taking the position that unless there is a clear constituency interest, their electoral mandate is to act as they themselves deem best. Others follow other strategies. The important point is whether any information is available that provides some fix on constituency opinion in general.

In short, we do not need to be naive and imagine that if constituency opinion were known with regard to this or that policy matter, representatives would scurry to respond in exactly that way. We know perfectly well that typically they will not, although some few innocents will try, and most, if not all the rest, will at least store the information as one among many forces to be reckoned with in their behavior. Nor need we be naive and imagine that most of the time constituency opinion as registered through direct questions of proper samples of constituents will differ markedly from what the representative is likely to learn more naturally from friends, the local editor, or party headquarters. Although the details are unknown, we know that many times there will be no palpable discrepancy at all. Perhaps such occasions are downright infrequent.

However, to the degree that representatives are prepared to respond to constituency opinion at all (and most are willing to pay at least faint attention because of the need for reelection), then the constituency opinion that they perceive might as well be the right one. It is in this connection that election research, with its detailed probing of the popular mandate, and the polls, with their more frequent updating of trends in opinion, can claim some social utility. However imperfectly they may do it, they do serve to restore a crucial link in the communication system that is the democratic process.

REFERENCES

Arrow, Kenneth
1958 Social Choice and Individual Values. New York: Wiley.

Berelson, Bernard R., Paul F. Lazarsfeld, and William N. McPhee
1954 Voting. Chicago: University of Chicago Press.

Campbell, Angus
1960 "Surge and decline: A study of electoral change." Public Opinion Quarterly 24(Fall):397-418.

Campbell, Angus, and Robert L. Kahn
1952 The People Elect a President. Ann Arbor, Mich.: Survey Research Center.

Campbell, Angus, Gerald Gurin, and Warren E. Miller
1954 The Voter Decides. Evanston, Ill.: Row, Peterson.

Campbell, Angus, Philip E. Converse, Warren E. Miller, and Donald E. Stokes

1960 The American Voter. New York: Wiley.

Condorcet, M. J. de

1785 Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. Paris: Imprimerie Royale.

Davis, Otto, Melvin Hinich, and Peter Ordeshook

1970 "An expository development of a mathematical model of the electoral process." American Political Science Review 64 (June):426-488.

Downs, Anthony

1957 An Economic Theory of Democracy. New York: Harper.

Gardner, Martin

1980 Scientific American (October):16-26.

Hatelling, Harold

1929 "Stability in competition." The Economic Journal 34:41-57.

Jacobson, Gary C.

1980 Money in Congressional Elections. New Haven, Conn.: Yale University Press.

Jones, Ruth S.

no date "State election campaigning in 1980." Unpublished paper.

Key, V. O.

1955 "A theory of critical elections." Journal of Politics 17:3-18.

Key, V. O.

1966 The Responsible Electorate. Cambridge, Mass.: Belknap-Harvard Press.

Kramer, Gerald

1971 "A decision-theoretic analysis of a problem in political campaigning." In Joseph Bernd, ed., Mathematical Applications in Political Science II. Dallas, Tex.: Arnold Foundation, SMU Press.

Kramer, Gerald

1971 "Short-term fluctuations in U.S. voting behavior, 1896-1964." American Political Science Review 65 (March).

Lazarsfeld, Paul F., Bernard R. Berleson, and Hazel Ga

1944 The People's Choice. New York: Duell, Sloan and Pearce.

Merriam, Charles E., and Harold Gosnell

1924 Non-Voting. Chicago: University of Chicago Press.

Miller, Warren
1964 "Majority rule and the representative system of government." Pp. 343-376 in Erik Allardt and Y. Littunen, eds., Cleavages, Ideologies and Party Systems: Contributions to Comparative Political Sociology. Helsinki: Transactions of the Wewtevmarck Society.

Miller, Warren E., and J. Merrill Shanks
no date "Policy directions and presidential leadership: alternative interpretations of the 1980 presidential election." British Journal of Political Science, forthcoming.

Miller, Warren E., and Donald E. Stokes
1963 "Constituency influence in Congress." American Political Science Review 57:47-56.

Mosteller, Frederick, Herbert Hyman, Philip J. McCarthy, Eli S. Marks, and David B. Truman
1949 The Pre-Election Polls of 1948: Report to the Committee on Analysis of Pre-Election Polls and Forecasts. New York: Social Science Research Council.

Nie, Norman, Sidney Verba, and John Petrocik
1976 The Changing American Voter. Cambridge, Mass.: Harvard University Press.

Pitkin, Hannah
1967 The Concept of Representation. Berkeley: University of California Press.

Pool, Ithiel, Robert Abelson, and Samuel Popkin
1964 Candidates, Issues and Strategies. Cambridge, Mass.: MIT Press.

Rossi, Peter
1959 "Four landmarks in voting research." In E. Burdick and A. Brodbeck, eds., American Voting Behavior. Glencoe, Ill.: The Free Press.

Siegfried, André
1913 Tableau Politique de la France de l'uest. Paris: Colin.

Smithies, Arthur
1941 "Optimum location in spatial competition." Journal of Political Economy 49:423-439.

Wahlke, John C., Heinz Eulau, William Buchanan, and Leroy Ferguson
1962 The Legislative System: Explorations in Legislative Behavior. New York: Wiley.

Wolfinger, Raymond E., and Steven J. Rosenstone
1980 Who Votes? New Haven: Yale University Press.

Behavior and Health: The Biobehavioral Paradigm

David S. Krantz, David C. Glass,
Richard Contrada, and Neal E. Miller

INTRODUCTION

The social and behavioral sciences have become increasingly significant for problems of physical health. These disciplines have matured and expanded beyond the area of mental health to a far broader one called behavioral medicine, which is concerned with behavioral factors in physical disease. At the same time, old disciplinary boundaries are being erased; behavioral and biomedical scientists alike are studying the joint influence of psychosocial and biological factors on somatic health and illness.

Disease has been viewed by Western medicine as a biological phenomenon, a product of specific agents or pathogens and bodily dysfunction. However, this biomedical model has not accounted for all states of illness, nor has it explained selective susceptibility and the fact that certain diseases occur in some people but not in others. The need for a broader model of health and illness, encompassing psychological and social variables and their interaction with biological processes, has been recognized by both the biomedical community (Engel, 1977) and by behavioral scientists (e.g., Miller, 1976; Matarazzo, 1980). Many medical problems, including some of the most common in modern society (heart disease, cancer), appear to be influenced by behavioral and social variables, such as habits of living (smoking, diet, exercise) or by what

---

This paper was commissioned by the Social Science Research Council for The National Science Foundation's Five Year Outlook on Science and Technology: 1981.

has been termed psychosocial stress. For example, epidemiologists studying lung cancer have considered many causal variables, including heredity and the physical environment (particularly air pollution), but the strongest risk factor turns out to be the behavioral variable of cigarette smoking.

In the United States at the turn of the century the greatest contributors to morbidity and mortality were infectious diseases. Today, the leading causes of mortality are chronic diseases, including cardiovascular disorders and cancers. These disease states are caused by a confluence of social, environmental, behavioral, and biological factors (Institute of Medicine, 1978; U.S. Department of Health, Education, and Welfare, 1979a). The crucial role of behavioral variables in today's most pressing health problems is clearly stated in the 1979 Surgeon General's report on health promotion and disease prevention: " . . . of the ten leading causes of death in the United States, at least seven could be substantially reduced if persons at risk improved just five habits: diet, smoking, lack of exercise, alcohol abuse, and use of antihypertensive medication" (U.S. Department of Health, Education, and Welfare, 1979a:14).

In recent years, important associations between psychosocial variables and physical disease outcomes have been documented. A biobehavioral paradigm has emerged from these efforts to advance scientific understanding beyond the descriptive level. In contrast to a purely correlational approach, biobehavioral research explores basic mechanisms linking behavioral processes to disease states. This research involves the integration of behavioral science principles and methods with biomedical knowledge of the disease being studied. An example is provided by recent evidence linking such psychosocial factors as emotional stress to the development of cardiovascular disease. Behavioral scientists working in this area are devoting increasing attention to the physiological processes (e.g., neuroendocrine activity) implicated in the development of coronary heart disease in order to determine how these processes are influenced by behavioral events.

The processes linking behavior to physical illness of various kinds may be grouped into three broad categories: direct psychophysiological effects, health-impairing habits and life-styles, and reactions to illness and the sick role.

Direct Psychophysiological Effects

The first category involves alterations in tissue function via neuroendocrine and other physiological responses to psychosocial stimuli. This mechanism encompasses bodily changes without the intervention of external agents, such as cigarette smoking or dietary risk factors, although the two sets of variables may produce interactive effects (e.g., stress and smoking might increase, synergistically, the risk of coronary heart disease). Central to this mechanism is the concept of stress, which was originally described by Hans Selye (1956) as a nonspecific response of the body to external demands that are placed on it. According to Selye, the stress response proceeds in a characteristic three-stage pattern, which involves a variety of physiological systems (neural, hormonal, and metabolic) in complex interrelation with each other. The term stress is also used in a psychological sense (Cox, 1978; Lazarus, 1966) to refer to the internal state of an individual who is perceiving threats to physical and/or psychic well-being. This broader use of the term places emphasis on the organism's perception and evaluation of potentially harmful stimuli and considers the perception of threat to arise from a comparison between the demands imposed on the individual and his or her felt ability to cope with these demands. A perceived imbalance in this mechanism gives rise to the experience of stress and to the stress response, which may be physiological and/or behavioral in nature.

Physiological responses to stress include neural and endocrine activity, which in turn can influence a wide range of bodily processes, including metabolic rate, cardiovascular and autonomic nervous system functioning, and altered immune reactions (Levi, 1979; Mason, 1971). Short-term stress responses include hormonal and cardiovascular reactions (e.g., increased heart rate and higher blood pressure), which may precipitate such clinical disorders as stroke, cardiac instabilities and pain syndromes, and psychosomatic symptoms in predisposed individuals. If stimulation becomes pronounced, prolonged, or repetitive, the result may be chronic dysfunction in one or more systems (e.g., gastrointestinal, cardiovascular, etc.).

Early stress research (Selye, 1956) emphasized the generality or nonspecificity of responses to a wide variety of stimuli, but subsequent work has recognized that the link between stress and disease is not simple;

instead, it depends on the context in which the stressful agent occurs, how individuals appraise it, and the social supports and personal resources available (Lazarus, 1966; F. Cohen et al., 1980; Mason, 1971). There are wide individual differences in physiological responses to stressors, which depend not only on biological predispositions (Levi, 1979), but also on the individual's felt ability to cope with or master conditions of harm, threat, or challenge. For example, stressful events are inevitable throughout the life cycle, yet few individuals suffer lasting adverse effects. Research has shown that various social and psychological factors (e.g., styles of coping or the social supports provided by others) act to modify or buffer the impact of stressful events on illness (F. Cohen et al., 1980).

It should be emphasized that since direct psychophysiological effects involve functional alterations brought about in part by exposure to psychosocial stimuli, adequate scientific understanding requires a specification of mediating physiological processes (e.g., the neuroendocrine and hormonal systems). It is not enough to establish correlations between disease end points and behavioral variables.

Health-Impairing Habits and Life-Styles

Second, behavior can lead to physical illness when individuals engage in habits and styles of life that are damaging to health. Personal habits play a critical role in the development of many serious diseases, as amply documented by the recent Surgeon General's reports on smoking and health and health promotion and disease prevention. Cigarette smoking is probably the most salient behavior in this category, for it has been implicated as a risk factor for three leading causes of death in the United States--coronary heart disease, cancer, and stroke. However, poor diet, lack of exercise, excessive alcohol consumption, and poor hygienic practices also have been linked to disease outcomes. These habits may be deeply rooted in cultural practices or initiated by social influences (e.g., smoking to obtain peer group approval). They may be maintained as part of an achievement-oriented life-style as well as by the interaction of biological and behavioral mechanisms of addiction. Therefore, a major focus of research in behavioral medicine has been the role of sociocultural systems, life-styles, and psychophysiological processes in the etiology and pathogenesis of chronic diseases. Considerable attention also has been directed toward the development of techniques to modify those behaviors that constitute risk factors for illnesses.

Reactions to Illness and the Sick Role

Third, behavior can lead to physical illness when individuals minimize the significance of symptoms, delay seeking medical care, or fail to comply with treatment and rehabilitation regimens. One prominent example is the sizable number of heart attack patients who procrastinate in seeking help, thereby endangering their chances of survival. These actions are representative of a larger area of study concerned with the way people react to the experience of organ dysfunction (illness behavior) as well as to the experience of being a sick person (patienthood). To succeed, medical therapy requires that the patient follow the physician's advice, but an extensive literature reports disturbingly low rates of compliance with health and medical care regimens (Sackett and Haynes, 1976). Accordingly, there has been considerable research on social and psychological processes involved in patients' reactions to pain and illness, the decision to seek medical care, and medical compliance. This research has led to the development of interventions that have been applied in treatment and rehabilitation settings.

Overview

The categories outlined above direct attention to the range of behavioral variables acknowledged as important factors in somatic health and illness. Traditionally, biomedical and behavioral scientists have studied many of these same problems independently of one another and from different perspectives. In recent years, there has been more interdisciplinary contact as well as a growing audience for complex problems of behavior and health.

In a paper of this length it is impossible to represent fully the broad spectrum of health-related behavioral research. We therefore do not attempt to discuss mental illness or substance abuse disorders, except insofar as they are related to physical disease end points. Instead we consider first the literature on behavioral and social factors in the etiology and pathogenesis of selected physical diseases such as cancers, psychosomatic disorders such as ulcers, and infectious disease. Because considerable progress has been made in understanding the relationship between behavior and the major cardiovascular disorders (the leading cause of death in the United States), emphasis is directed to this area as an exemplar of the biobehavioral approach to physical illness. Next we consider rehabilitation and the prevention of physical disease. This discussion is not limited to any single disorder but instead emphasizes themes that are relevant to a variety of somatic illnesses. We conclude with an overview of research directions projected for the next five years.

BEHAVIORAL FACTORS IN THE ETIOLOGY AND PATHOGENESIS OF PHYSICAL DISEASE

In 1900 the leading causes of death in the United States were pneumonia, influenza, and tuberculosis. Changing patterns of illness since that time have been marked by the ascendancy of cardiovascular diseases as the chief causes of mortality in this country. The cardiovascular disorders, including coronary heart disease and high blood pressure, now account for more than half of all deaths. A large percentage of these would be classified as premature, for they occur during the middle years, ages 35 to 50 (National Science Foundation, 1980).

Atherosclerosis, Coronary Heart Disease, and Sudden Death

Coronary atherosclerosis is a symptomless condition characterized by narrowing and deterioration of the arteries, including the coronary arteries (that is, the blood vessels that nourish the heart). An excess accumulation of cholesterol and related lipids forms a mound of tissue, or atherosclerotic plaque, on the inner wall of one or more of the coronary arteries (Hurst et al., 1978). The formation of atherosclerotic plaque may proceed undetected for years, affecting cardiac functioning only when it causes a degree of obstruction sufficient to diminish blood supply to the heart. Once this occurs, coronary atherosclerosis has evolved into coronary heart disease (CHD).

In one form of CHD, angina pectoris, occasional instances of inadequate blood supply (ischemia) cause the individual to experience attacks of chest pain. Although ischemia per se does not cause permanent tissue damage, angina is a painful condition that can lead to more serious complications. A more severe and frequently fatal consequence of CHD is myocardial infarction, or heart attack, in which a prolonged state of ischemia results in the death of a portion of the heart tissue. Other manifestations of CHD include congestive heart failure, conditions secondary to myocardial infarction (ventricular failure, heart rupture), and disturbances of the conductive or beat-regulating portion of the heart, i.e., the arrhythmias (Hurst et al., 1978).

Standard Risk Factors for Cardiovascular Disease

Individuals who are likely to develop coronary heart disease may be identified with a modest degree of accuracy. This is possible because a set of risk factors, attributes of the population of interest or of the environment that appear to increase the likelihood of developing one or more of the clinical manifestations of cardiovascular disease, has been recognized in recent years. The following risk factors have been identified: (1) aging, (2) sex (being male), (3) elevated serum cholesterol and related low-density lipoproteins, (4) dietary intake of animal fats and cholesterol, (5) high blood pressure, (6) heavy cigarette smoking, (7) diabetes mellitus, (8) specific diseases such as hypothyroidism, (9) family history of coronary disease, (10) obesity, (11) sedentary life-style, and (12) specific anomalies of the electrocardiogram, such as evidence of left ventricular hypertrophy (Kannel et al., 1976).

It should be emphasized that the predictive value of these variables is in most cases far from being a settled issue. For example, the pathogenic influence of a sedentary life-style and obesity per se failed to receive empirical support in a number of investigations (Mann, 1974). Early emphasis on the role of dietary fat intake has also been challenged in recent years (Mann, 1977). Suffice it to say here that coronary disease has a multifaceted etiology that involves many of the factors listed above in varying degrees of importance.

The effects of these standard risk factors have been viewed in terms of their physiological influence (e.g., the toxic effects of tars and nicotine, the role of salt intake in regulating blood pressure levels, the relationship between diet and serum cholesterol); however, many of these variables are determined, at least in part, by behavioral factors. For example, cigarette smoking is a preventable behavior undoubtedly brought about by psychosocial forces (Leventhal and Cleary, 1980). Cultural, racial, and social class groups differ in serum cholesterol levels, independently of dietary practices (McDonough et al., 1965). Enhanced risk due to sex and age may derive from nonbiological correlates of these variables, such as occupational pressure, stressful life events, and behavior patterns (Eisdorfer and Wilkie, 1977; Riley and Hamburg, 1981).

Resting blood pressure levels differ among racial groups, socioeconomic status, and cultures (Weiner, 1977). A family history of coronary disease, while in some cases linked to a specific genetic mechanism (Goldstein and Brown, 1974), may also lead to enhanced risk of CHD through such psychosocial channels as family patterns of cigarette smoking, diet, and socioeconomic condition. At least two implications follow from these observations: Analysis of the etiology and pathogenesis of coronary disease must include the domain of psychosocial factors; and psychosocial factors should be considered important targets in the prevention and treatment of CHD.

Psychosocial Risk Factors for Coronary Heart Disease

The most predictive combinations of the standard risk factors fail to identify most new cases of coronary heart disease (Jenkins, 1971). Some variable or set of variables appears to be missing from the predictive equation. This limitation in knowledge has led to a broadened search for influences and mechanisms contributing to coronary risk; it now includes social indicators such as socioeconomic status and social mobility, and psychological factors such as anxiety and neuroticism, psychological stress, and overt patterns of behavior. The results have been encouraging, though not uniformly so. The two most promising psychosocial risk factors to emerge in recent years are psychological stress and the Type A coronary-prone behavior pattern (Jenkins, 1971).

Psychological Stress As noted earlier, Selye (1956) first popularized the notion of stress, which he defined as the body's nonspecific physiological reaction to noxious agents or stressors. More recently, psychological investigators such as Lazarus (1966) and Mason (1971) have taken exception to this view, arguing that the body's response varies with the particular type of stressor and the context in which the stressor occurs (Lazarus, 1966; Glass and Singer, 1972).

Several indices of psychological stress have been studied in relation to the development of coronary heart disease. Research suggests that excessive work and job responsibility may enhance coronary risk, especially when they approach the limits of the individual's capacity to control the work environment (Haynes et al., 1980; House, 1975). Another job-related stressor that appears to be related to coronary disease is reported work dissatisfaction, such as lack of recognition by superiors, poor relations with coworkers, and inferior work conditions (House, 1975). Other dissatisfactions, including problems and conflicts with finances and family, have been correlated with the presence and future development of coronary heart disease (Haynes et al., 1980; Medalie et al., 1973).

The experience of a single, traumatic life event has long been suspected as a cause of clinical CHD (Cannon, 1942). More recently, it has been suggested that the cumulative effects of repeated adjustments required by life changes drain the adaptive resources of the individual and increase susceptibility to a variety of diseases. To test this, an objective instrument, the Social Readjustment Rating Scale (SRRS), was developed by Holmes and Rahe to assess the impact of such events as the death of spouse, a change to a different line of work, and a son's or daughter's leaving home (Holmes and Rahe, 1967).

Several retrospective studies have used this technique in an effort to link the accumulation of life events with the occurrence of coronary heart disease (see Garrity and Marx, 1979). For example, survivors of myocardial infarction show a pattern of increased life changes during the previous period of approximately one and one-half years, whereas healthy control subjects reported a relatively stable number of life events during the same period. Other research, in which information regarding life events prior to sudden cardiac death was obtained from a survivor of the deceased (usually the spouse), revealed an accumulation in the intensity of life events in the six months prior to death (Garrity and Marx, 1979).

Despite replication of these findings, negative results have been reported as well (e.g., Hinkle, 1974). Reviewers point to defects in the methodology of retrospective designs that might account for the positive findings (Dohrenwend and Dohrenwend, 1978). However, such explanations cannot explain significant associations obtained in prospective studies in which data concerning psychosocial stressors were obtained prior to the development of disease (e.g., Haynes et al., 1980; Medalie et al., 1973).

The relation of stress to pathological outcomes depends on both the adaptive capacity of the individual before the stress occurs and the resources marshalled in response (F. Cohen et al., 1980). It follows that variables moderating the impact of stress must be taken into account in order to gauge the predictive validity of stress as a risk factor for coronary disease. These moderators include biological factors (e.g., genetic susceptibility, general state of health); psychological attributes (e.g., felt ability to cope); aspects of the immediate context in which the stressor occurs (e.g., whether the stressor is perceived as controllable); various sociocultural variables (e.g., amount of social support from other people and/or the health care system); and factors related to the life course (e.g., the expectedness of events at a certain stage of life). For example, there is evidence that individuals who have social supports may live longer, have a lower incidence of somatic illness, and possess higher morale and more positive mental health (F. Cohen et al., 1981).

Type A Coronary-Prone Behavior Pattern Perhaps the most thoroughly investigated psychosocial risk factor for coronary disease is the Type A behavior pattern (Rosenman and Friedman, 1974). Type A (or Pattern A) is characterized by extreme competitiveness and achievement striving; a strong sense of time urgency and impatience; hostility; and aggressiveness. The relative absence of these traits is designated as Type B behavior.

The Type A concept does not refer simply to the conditions that elicit its characteristic behavior, or to the responses per se, or to some hypothetical personality trait that produces them. It refers instead to a set of behaviors that occur in susceptible individuals in appropriately stressful and/or challenging conditions. Type A is therefore the outcome of the interaction of a person and a situation. It is not a typology but a behavior pattern, which is displayed in varying degrees at one time or another by everyone.

People who consistently display Type A characteristics have long been suspected of being at greater risk for clinical CHD (Osler, 1892). However, the major impetus for research validating this hypothesis comes from work initiated by cardiologists Meyer Friedman and Ray Rosenman only two decades ago. These investigators developed a structured interview that constitutes the major tool for the diagnosis of Pattern A. More recently, two self-administered questionnaires have been developed to detect Type A behavior.

Although several studies have documented an association between Pattern A and CHD, the most convincing evidence comes from the Western Collaborative Group Study (Rosenman et al., 1975). In this prospective, double-blind study, more than 3,000 initially healthy men ages 39-59 were assessed for a comprehensive array of social, dietary, biochemical, clinical, and behavioral variables. A follow-up for eight-and-a-half years showed that subjects exhibiting Type A behavior at the study's inception were about twice as likely as Type B individuals to develop angina pectoris or myocardial infarction. This twofold differential in risk remained when statistical procedures were used to control for the influence of other risk factors such as cigarette smoking, serum cholesterol, and high blood pressure. This research also has linked Pattern A to sudden cardiac death (Friedman et al., 1973) and recurrent myocardial infarction.

A follow-up for eight years of data from the Framingham study, a large-scale prospective study of heart disease undertaken by the National Insitutes of Health, indicated that Pattern A is predictive of CHD in both men and women, although for men the enhanced risk appeared only among white-collar workers (Haynes et al., 1980). After controlling for the influence of the traditional risk factors, it was found again that Pattern A conferred increased CHD risk. The prospective association of Pattern A with coronary disease in the Western Collaborative Group Study and the Framingham study constitutes strong evidence for the independent pathogenic influence of the behavior pattern.

There may also be an association between Pattern A and coronary atherosclerosis. Supporting evidence has been obtained through the use of coronary angiographic techniques that make it possible to quantify the extent of coronary artery disease in living patients (Blumenthal et al., 1978). In a recent study of men who underwent repeated coronary angiograms, an association was found between Pattern A and the progression of atherosclerosis (Krantz et al., 1979). It should be noted, however, that evidence for the association between Pattern A and coronary atherosclerosis is not unequivocal (Dimsdale et al., 1980).

Pathophysiological Mechanisms Linking Stress and Behavior Pattern A to Coronary Disease It is not enough to demonstrate a relationship between CHD risk factors--whether biomedical or psychosocial in nature--and the occurrence of cardiovascular disease. The precise mechanisms mediating the association must be specified. Although the pathogenesis of coronary heart disease is not completely understood, several factors are believed to play a major contributing role. These include a variety of physiological and biochemical states that may enhance coronary risk by influencing the initiation and progression of atherosclerosis and/or by precipitating clinical CHD (Herd, 1978; Ross and Glomset, 1976). Many of these physiological states have been observed in experimental studies of psychological stress. For example, hemodynamic effects, such as elevated heart rate and blood pressure, and biochemical changes, such as increased levels of serum cholesterol, are produced in animals under prolonged or severe stress (Schneiderman, 1978). It has been observed, in addition, that a reduction in blood clotting time occurs under conditions of stress, and in some cases degeneration of heart tissue has been reported as well. Other animal research has linked laboratory stressors to a lowered threshold for ventricular arrhythmia and for ventricular fibrillation (e.g., Lown et al., 1973), a state that leads to sudden cardiac death unless immediate treatment is given.

Potentially pathogenic states have been observed in studies of psychological stress in healthy humans. For example, life stressors such as occupational pressure have been shown to produce biochemical changes such as elevated levels of serum cholesterol (Friedman et al., 1958). Other research has demonstrated an association of increased heart rate and blood pressure with stressors such as the performance of mental arithmetic, harassment, and the threat of electric shock. Still other studies report that the stresses of automobile driving, public speaking, and discussion of emotionally charged topics provoke ventricular arrhythmia (Herd, 1978).

A notable feature of the foregoing research is the measurement of physiological reactivity in response to stress, as distinct from the observation of basal or resting levels of physiological variables. These changes in functioning, which are not detected by basal measurement of risk factors, are believed to yield a better index of the pathogenic processes involved in coronary heart disease. In addition, by observing such changes in response to real-life or laboratory-induced stressors, pathogenic states may be detected within the context of their psychosocial antecedents.

The physiological concomitants of psychological stress are believed to result from the activation of the sympathetic-adrenal medullary system (SAM) and the pituitary-adrenocortical axis (PAC). Interest in the impact of SAM activation on bodily reactions to emergency situations may be traced to Walter Cannon's work on the fight or flight response. This neuroendocrine response appears to be elicited in situations demanding effortful coping with threatening stimuli (Frankenhaeuser, 1971). The hormonal responses of the PAC axis were emphasized by Selye in his notion of a generalized physiological response to aversive stimulation. The PAC secretions include a number of hormones that influence bodily systems of relevance to the development of coronary disease. The corticosteroids, which include cortisol, regulate the metabolism of cholesterol and other lipids involved in the atherosclerotic process. Activation of the SAM system also may have a special significance in mediating stress-related pathophysiological changes. Particularly culpable in this regard is secretion of the catecholamines, epinephrine and norepinephrine, which are believed to induce many of the pathogenic states associated with psychological stress. These include increased blood pressure and heart rate, elevation of blood lipids, acceleration of the rate of damage to the inner layers of the coronary arteries over time, and provocation of ventricular arrhythmias, believed to lead to sudden death.

The same pathophysiological mechanisms linking stress and coronary heart disease may apply, a fortiori, to Type A individuals, thereby accounting in part for their enhanced coronary risk. Research has shown greater urinary catecholamine secretion during the working day and greater plasma catecholamine responses to competition and stress among Type A individuals compared with Type B individuals

(Rosenman and Friedman, 1974). Recent studies by Glass et al. (1980) have also demonstrated higher elevations in plasma catecholamines among Type A individuals in situations of hostile competition. Subjects were led to believe they would compete for a prize with an opponent (who was actually an experimental confederate) on a challenging electronic game. A and B subjects were exposed to one of two experimental conditions. In the no-harass condition, the confederate remained silent and simply competed against the subject. In the harass condition, the same procedure was followed with one variation: The confederate made a series of remarks designed to harass the subject. Throughout the task, blood pressure and heart rate were measured and blood samples were drawn via an in-dwelling venous catheter. Results showed that harassment had an effect on plasma catecholamines, blood pressure, and heart rate for both types of subjects. However, of those harassed, Type A individuals showed increased elevations of plasma epinephrine, systolic blood pressure, and heart rate. These cardiovascular and neuroendocrine changes are consistent with the findings of other investigators indicating greater cardiovascular reactivity among Type A than Type B individuals in challenging situations (Dembroski et al., 1978; Herd, 1978).

High Blood Pressure

High blood pressure (also called essential hypertension) is a condition of unclear etiology in which blood pressure shows chronic elevations. When the disorder becomes developed fully, increased pressure is usually due to a constriction or contraction of blood vessels throughout the body (Page and McCubbin, 1966). Although high blood pressure is a symptomless disorder, there is epidemiological evidence that even mild blood pressure elevations are associated with a shortening of life expectancy (Kannel and Dawber, 1971) and increased risk of coronary heart disease and stroke. As is the case with coronary heart disease, the causes of high blood pressure are believed to involve complex interactions between genetic, sociocultural, behavioral, and physiological processes.

Heterogeneity of the Disorder and Physiological Mechanisms

Essential hypertension is not a single, homogeneous disease. In the development of the disorder, blood pressure is thought to progress over a period of years from moderately elevated or borderline levels to more appreciably elevated levels, called established hypertension. Several pathogenic mechanisms may bring about blood pressure elevations, and different physiological and/or behavioral mechanisms are implicated at various stages of the disorder. For example, individuals with borderline hypertension are commonly observed to have an elevated cardiac output (i.e., amount of blood pumped by the heart) but show little evidence of increased resistance to the flow of blood in the body's vasculature (Julius and Esler, 1975). As noted earlier, this physiological pattern is consistent with increased activation of the sympathetic nervous system, which is the body's initial reaction to psychological stress. However, in older individuals with more established high blood pressure, cardiac output is either normal or depressed, while the vascular resistance is elevated.

Although psychological stimuli such as emotionally stressful events have been shown to correlate highly with the exacerbation of hypertensive episodes in diagnosed patients (Weiner, 1977), recent research on behavioral influences has focused increasingly on earlier stages, rather than on the culmination of the disease. In addition to cardiovascular adjustments and changes, the physiological mechanisms of high blood pressure probably involve the interaction of the central and autonomic nervous systems, the endocrine-hormonal system, and the kidneys. Accordingly, behavioral factors (in particular, psychological stress) may play a role in the etiology of high blood pressure via a number of physiological pathways (Kaplan, 1980). Recall that stress leads to discharge of the sympathetic nervous system and to increases in catecholamines. High levels of blood and tissue catecholamines have been found in some hypertensive humans and animals (Julius and Esler, 1975). Such elevations could lead to increased blood pressure via increased heart rate and force of heart action, constriction of peripheral blood vessels, and/or activation of a hormonal mechanism in the kidney that constricts the vasculature (Kaplan, 1980) and regulates the volume of blood (see below). It should be noted that the cardiovascular system has an intrinsic means of regulating blood flow--namely, the constriction of blood vessels whenever blood flow is increased (autoregulation). Elevated output of blood from the heart produced by nervous system activity may also lead to pressure elevations via this mechanism. There has been much investigation of the role of the kidney in hypertension. The enzyme renin, which is released by the kidney, is involved in a physiological regulatory process (the renin-angiotensin-aldosterone mechanism), which leads the kidney to increase water reabsorption and expand the volume of blood, thus raising the pressure (Kaplan, 1980). The process of renin release normally should be dampened whenever the blood pressure is raised, but in a subcategory of individuals with high blood pressure, the level of renin in the blood is inappropriately high. Thus, presumably, either with or without the involvement of the sympathetic nervous system, the renin-angiotensin system is an important mechanism in the etiology and maintenance of high blood pressure.

Genetic-Environment Interactions

The prevalence of essential hypertension in the United States usually increases with age, and below the age of 50 years it occurs with less frequency in women than in men (Weiner, 1977). Evidence from animal research and studies of human twins indicates that genetic factors play a role in the etiology of the disease (Pickering, 1967). This evidence suggests that many genes are involved in the susceptibility to high blood pressure, and it is likely that in humans sustained elevations in blood pressure are produced by an interaction of a variety of environmental and genetic factors. Consider that epidemiological studies reveal a difference in the prevalence of high blood pressure among various social and cultural groups, a difference that cannot be accounted for by genetic factors alone (Henry and Cassel, 1969). For example, in the United States hypertension is more common among blacks than among whites, and the prevalence of high blood pressure is greater in poor than in middle-class black Americans (Harburg et al., 1973). Animal research, described below, similarly reveals examples of environmental factors (such as dietary salt intake or environmental stress) that lead to sustained blood pressure elevations only in certain genetic strains. Closely related to this observation is the finding that family members tend to have similar blood pressures. Although the prevalent view attributes this solely to a genetic source, there is emerging evidence suggesting joint genetic and environmental effects. A possible environmentally determined behavioral factor, family social interaction, is illustrated by a recent study in which more negative nonverbal behavior (e.g., grimacing, gaze aversion) was observed among families with a hypertensive father, than in families with a normotensive father (Baer et al., 1980).

Behavioral Factors

Sociocultural and psychological studies of humans, in conjunction with animal research, have identified some environmental factors related to behavior that might play a role in the initiation of high blood pressure. These factors include dietary intake of salt, obesity, and psychological stress.

Dietary Salt Intake Much has been written about the role of salt in essential hypertension, largely because excessive intake of sodium is thought to increase the volume of blood. However, studies of the relationship between salt intake and high blood pressure indicate that high salt intake may result in sustained blood pressure elevations only in genetically predisposed individuals. For example, Dahl et al. (1962) found that there are salt-sensitive and salt-resistant strains of rats; more recently, studies of increased sodium intake in humans have noted that mild hypertensive patients can be divided into groups that are differentially sensitive to increased salt intake (Kawasaki et al., 1978).

Research also suggests that reduction in sodium intake can result in decreased blood pressure (Shapiro, 1982). Evidence that decreasing sodium intake in the diet will lower blood pressure in humans derives from several sources: from studies showing the effectiveness of diuretic (sodium- and fluid-excreting) medication; from studies demonstrating that drastic sodium restriction can measurably lower blood pressure (Shapiro, 1982); and, conversely, from studies indicating that healthy persons put on high sodium diets show pressure increases. At present we do not fully understand the human craving for salt intake in excess of physiological needs, but evidence suggests that it may in part be a habit that is learned.

Obesity Obesity is another cultural and behavioral phenomenon that plays an important role in hypertension, although the precise reasons for the higher prevalence of high blood pressure in obese patients have not been determined. In the case of obesity, recent studies have determined that weight loss can result in significant decreases in blood pressure (Shapiro, 1982). As is the case with all nonpharmacological approaches involving life-style alterations, effective treatment outcomes depend not only on producing transitory changes in behavior but also on the maintenance of these changes and sustained compliance with prescribed regimens. We will discuss these issues more fully in a later section of this paper.

Psychosocial Stress Stress deriving from psychosocial causes is yet another factor implicated in the etiology and maintenance of high blood pressure. As previously described, psychological stimuli that threaten the organism result in cardiovascular and endocrine responses that can play an important role in the development of hypertension (Julius and Esler, 1975). The brain and the central nervous system, which are involved in determining whether situations are harmful or threatening, play a role in physiological mechanisms mediating the impact of noxious stimuli. On a societal level, there is evidence that blood pressure elevations occur under conditions of rapid cultural changes and socioeconomic mobility. Moreover, there are many studies in which primitive populations living in small, cohesive societies were found to have low blood pressure that did not increase with age. When members of such societies migrated to areas in which they were suddenly exposed to Western culture, they were found to have high levels of blood pressure that increased with age. This suggested some cumulative effect of the new living conditions that became evident over the course of the life-span (Henry and Cassel, 1969).

While such studies can attempt to rule out confounding factors, (e.g., diet, sanitation, etc.) by using carefully matched control groups and by employing statistical control techniques, there are inherent limits to conclusions that can be reached from correlational research. Experimental techniques for inducing high blood pressure in animals offer the ability to control both genetic and environmental factors by manipulating separate variables relevant to the course of this disorder (Campbell and Henry, 1982). Accordingly, various animal models of experimental high blood pressure indicate that the brain participates at some stage in the development or maintenance of increased blood pressure levels. The role of stress in the etiology of hypertension is supported by experimental studies demonstrating that sustained and chronic blood pressure elevations can be produced in animals exposed to environmental events such as fear of shock, social isolation followed by crowding, and experimentally produced conflict (see Campbell and Henry, 1982). Studies have also shown that animals placed on learning schedules that reward them for conditions in which blood pressure remains elevated display sustained blood pressure elevations. Thus, there is the possibility that learning and conditioning processes may be involved in the development of the hypertensive state. In accord with our earlier discussion of genetic-environment interaction, studies have demonstrated that strains of animals genetically susceptible to hypertension are also susceptible to stress-induced pressure elevations. Friedman and Dahl (1975) identified a genetic strain of rats that develops severe hypertension if excess salt is ingested. While on a low salt diet these rats were subjected to an experimental treatment (called approach-avoidance conflict) that involved punishment for bar-pressing responses necessary to obtain food. Other rats of the same genetically susceptible strain were yoked to these rats and received the same amount of food and shock but were not subjected to conflict; still others served as controls, receiving no shock but having free access to food. Results indicated that those rats exposed to the conflict of punishment for eating generally exhibited the highest mean blood pressures, followed by those given shock without conflict and also deprived of food. Furthermore, in a related study Friedman and Iwai (1976) showed that a genetic strain not susceptible to salt-induced hypertension did not develop pressure elevations when subjected to this same food-shock conflict situation.

Associations between emotional and behavioral stimuli and the development and/or maintenance of high blood pressure receive additional support from human studies indicating that techniques such as biofeedback and relaxation training can be used to modify the stress-induced components of high blood pressure (Shapiro, 1982). These techniques, discussed in more detail in a later section, are designed to counteract pressure-increasing stimuli that operate through the central and the autonomic nervous systems.

Personality Correlates The traditional psychosomatic approach to high blood pressure proposed that one's emotional disposition or personality traits play a causal role in the development of chronic blood pressure elevations (Harrel, 1980). The individual susceptible to high blood pressure has been described as one with inhibited and poorly expressed anger (supressed hostility), and it has been suggested that inhibited anger results in stimulation of the autonomic nervous system leading to acute and eventually chronic high blood pressure. However, although emotional states such as anger do lead to cardiovascular adjustments resembling high blood pressure, and although identifiable traits have been observed in patients with high blood pressure (Weiner, 1977), on balance the search for a hypertensive personality has not yielded conclusive results. Patients with high blood pressure are not homogeneous in terms of either physiological or psychological characteristics. A recent study showed convincingly that 30 percent of a sample of young male patients with mild pressure elevations and high plasma renin levels displayed both elevations in sympathetic nervous system activity and higher levels of suppressed hostility (Esler et al., 1977), a behavioral trait independently linked to increased nervous system activity. Explanations for these results might include suppressed hostility as leading to blood pressure elevations, by increased nervous system activity as the initial event, or by some other underlying factor. This issue may be resolved with studies of families of patients with high blood pressure and the social interaction patterns that could have a bearing on the personality development of offspring, or by research that looks at behavioral characteristics associated with the development of high blood pressure in susceptible strains.

Behavioral-Cardiac Interactions

Experimental work has sought to identify individuals who are at risk of developing high blood pressure and the types of situations that might activate genetic predispositions to the disorder. Since the aim is to understand mechanisms in the cause of the disorder, research has focused increasingly on the beginning stage rather than the culmination of the disease. Given that borderline high blood pressure is characterized by heightened responsiveness of the cardiovascular and sympathetic nervous systems to psychological stimuli such as mental stress (Julius and Esler, 1975), recent research has examined the tendency toward large episodic or acute increases in heart rate, blood pressure, and hormonal (catecholamine) activity of the sympathetic nervous system as possible mechanisms involved in etiology. Several groups of investigators (Manuck and Schaefer, 1978; Obrist, 1981) have found that cardiovascular responsiveness is a stable and persistently evoked response that can be measured reliably in a laboratory situation. Cardiovascular responsiveness to certain psychological stimuli has also been related consistently to family history of high blood pressure (a hypertension risk factor), even among individuals who have normal resting blood pressure levels and display no overt signs of the disorder (Falkner et al., 1979; Obrist, 1981). For example, in one representative study (Falkner et al., 1979), adolescents with normal blood pressure and at least one parent with high blood pressure displayed greater diastolic blood pressure and heart rate and plasma catecholamine responses to a stressful mental arithmetic task than did a control group of adolescents with no family history of high blood pressure.

Obrist (1981) reports that cardiovascular responsiveness above the level that is efficient for the body's metabolic needs results from situations in which active coping or behavioral adjustments are required. In active coping situations the organism tries to exert some behavioral control over a stimulus. By contrast, sympathetically mediated cardiovascular responses do not seem to be elicited in similar intensity or kind for other stressors (e.g., a stressful film) for which the individual remains passive and does not take direct action to attempt to control the situation. This concept of active coping may underlie those psychosocial conditions (e.g., rapid cultural change, socioeconomic mobility) shown by epidemiological research to be associated with blood pressure elevations in some human populations (Henry and Cassel, 1969).

By what physiological mechanisms might active coping and the resulting periodic increases in sympathetic nervous system activity lead to chronic blood pressure elevations? Research employing sophisticated pharmacological manipulations (such as the administration of drugs that selectively block the action of certain neurons) reveals that active coping with stress can alter regulatory mechanisms involving two physiological processes. In one the heart pumps excessive amounts of blood, and in the other the kidney reabsorbs excessive amounts of sodium with a resultant increase in the volume of blood (Obrist, 1981). In a series of animal studies, stress decreased water excretion to an extent that exceeded the metabolic needs of the organism (Obrist, 1981) and exacerbated or precipitated high blood pressure in animals with impaired kidney function (Harrell, 1980). Studies of humans exposed to various stressors (e.g., Obrist, 1981) revealed that active coping is specifically associated with a pattern of cardiovascular response resulting in increased cardiac output. Further research employing selective blocking drugs with humans confirmed that these cardiovascular changes resulted from increased activity of the sympathetic nervous system.

In sum, these studies represent several approaches currently being used by social and behavioral scientists to understand the etiology of high blood pressure. They represent an attempt to move from the symptom-oriented and purely descriptive level to a focus on mechanisms. These early findings justify further experimental, longitudinal, and naturalistic studies.

Psychosomatic Diseases: The Example of Peptic Ulcer

Although there is evidence that a variety of physical disorders are to some degree caused or exacerbated by psychological or emotional factors, the terms psychosomatic and psychophysiological refer to those physical conditions that appear to be initiated primarily by psychological factors (American Psychiatric Association, 1968). It should be noted that these conditions involve actual organ pathology, often due to activity of the autonomic nervous system initiated by psychosocial stimuli. Common examples of physical conditions that may be subsumed under this category include, but are not limited to, tension and migraine headache, ulcer, asthma, and rheumatoid arthritis. We discuss peptic ulcer as a representative disorder in this category.

A peptic ulcer is a lesion or sore in the lining of the stomach or the duodenum, the upper part of the small intestine that lies immediately below the stomach. A basic problem with the term peptic ulcer is imprecise definition, for it is used to include both gastric and duodenal ulcers. These two forms have certain common characteristics, but also significant differences. For example, duodenal ulcer is usually associated with an increase in gastric secretion of hydrochloric acid and the stomach enzyme pepsin, whereas gastric ulcer is not necessarily characterized in this manner. Indeed, clinicians view gastric and duodenal ulcers as separate disorders, associated with different predisposing and initiating mechanisms (Weiner, 1977). Thus, the general term peptic ulcer can be misleading, for it is not a single disease. Differences in anatomical location, natural history, pathophysiology, symptoms, and response to treatment produce considerable heterogeneity. Moreover, ulcers are sometimes "quiet" in the sense that they cause no discomfort and remain unnoticed. Thus the researcher investigating this disorder in symptomatic humans is not studying a population representative of all those who have the disease.

Physiological Mechanisms

Although new evidence about the physiological mechanisms involved in the pathogenesis of ulceration has developed in recent years, the disease processes are still poorly understood. It seems clear that the secretion of acid and pepsin by the stomach is important. Hydrochloric acid is continually being secreted into the stomach, even during sleep, and the secretion rate can be markedly increased by a variety of environmental stimuli, including the smell, taste, sight, or even thought of food (Weiner, 1977). Pepsin, produced in the stomach, is an important enzyme involved in the digestion of proteins. In about 50 percent of patients with active duodenal ulcer disease, elevated resting secretions of these two substances can be observed.

Release of pepsin and gastric acid are at least in part under the control of the central nervous system, particularly the vagus nerve, which increases stomach motility and secretion. However, these substances are also under the complex control of a variety of other hormones produced in the stomach, pancreas, and kidney (Weiner, 1977). Both stimulation and destruction of selective sites of the brain can produce lesions of the stomach and/or duodenum (Brooks, 1967). There is little question that central neural processes can be critical in the regulation of gastric secretion and the production of some type of ulceration in both animals and humans. Whether these processes are active in the development of chronic ulceration in humans is less clear, largely because we lack biomedical knowledge of the precise physiology of ulcer disease.

Animal Research

Gastric lesions can be produced in animals by a wide range of manipulations, including restraint and immobilization, stressful conditioning techniques, and conflict situations as well as food deprivation and painful sensory stimulation (Weiner, 1977). It should be noted that such lesions in animals are different in a number of ways from human ulcers, and therefore inferences to human disease from this work should be made with caution.

Studies with rats demonstrate that prior experience, particularly early in life, affects individual susceptibility to ulceration. For example, rats handled early in life were less likely to develop stomach lesions when immobilized than those who were not handled. However, rat pups separated prematurely from their mothers are more susceptible to gastric lesions (see Weiner, 1977). The age and genetic strain of an animal also affect its susceptibility to restraint-induced lesions. In addition, within a particular strain, susceptibility to ulcers seems to be related to individual variations in serum pepsinogen, a substance released under neural influences (Weiner, 1977) and converted to pepsin.

A series of experiments by Weiss (1972) employed behavioral techniques to demonstrate the role of psychological stress in the production of stomach lesions. These studies equated the strength, duration, and frequency of electric shock by using pairs of rats with electrodes on their tails wired together in series. One member of the pair was given a signal that enabled it to predict the onset of shock and determine when it was safe; the other animal received a signal at random. The animal able to predict shock occurrence developed considerably fewer stomach ulcers than the other animal. Furthermore, if the rat had available a simple coping response so that it could learn to control the shock, it developed fewer stomach lesions than its helpless partner, who received exactly the same shocks without the ability to control them. By contrast, if the coping response involved conflict, so that the rat had to take a brief shock to escape a longer one, the results were reversed.

These studies demonstrate the role of psychological variables (e.g., controllability of shock, conflict) as distinguished from the physical intensity of noxious stimulation (held constant for all animals in the research) in the experimental production of lesions. Moreover, there was evidence in this research that individual differences in the animals' behavioral responses to unpredictable shock were related to the development of lesions. Under conditions of uncertainty, those animals that made many efforts to cope with noxious stimulation by actively trying to avoid shocks were most ulcer-prone.

Human Studies

Much of the research investigating the relationship between psychological variables and peptic ulcer disease in humans is confounded by methodological problems. These include uncertain diagnostic criteria, difficulties in psychophysiological measurement of gastrointestinal function, problems in selection of appropriate study populations (e.g., hospitalized versus nonhospitalized patients, length of illness, etc.), and inappropriate control groups. Moreover, there have been few longitudinal studies in this area, so it is difficult to determine whether psychological variables precede or follow illness. Nevertheless, there is firm evidence from research with humans that psychological factors (including stress) bear at least tentative relationships to the regulation of gastric secretion and to the initiation and/or precipitation of ulcer symptomatology.

Classic studies of patients with abnormal openings to the stomach (gastric fistulae) reveal a remarkable covariation between stomach secretions and a variety of emotional states such as anger, resentment, and depression (Wolf and Wolff, 1947). The traditional psychosomatic approach (Alexander, 1950) proposed that ulcer patients were characterized by a specific set of traits--namely, conflicts over the need to be dependent--which led to neural activation and increased gastric secretions mediated by the vagus nerve. Evidence bearing on this hypothesis is inconsistent, but at least one prospective study of trainees for the military found that those who were hypersecretors of pepsinogen, an inherited risk factor for ulcers, could be identified on the basis of psychometric testing.

There have also been studies of individuals in stressful environments and occupations (Wolf et al., 1979). Recall that these situations are also associated with cardiovascular disorders, including coronary heart disease and high blood pressure. Cobb and Rose (1973) found that peptic ulcer was nearly twice as prevalent among air traffic controllers as among a matched control group in another occupation. In this study the prevalence of ulcers was higher among workers in high-stress control towers than among those in lower-stress control towers. In addition there is evidence, which has been confirmed repeatedly, of a higher prevalence of peptic ulcer disease in men who have supervisory jobs (i.e., foremen) than in executives and craftsmen (Wolf et al., 1979).

Predispositions to Specific Disorders

The field of psychosomatic research has provided medicine with a basis for predicting who might be at risk for a specific illness as well as knowledge of the conditions under which the predisposed individual is most likely to develop that disorder. At one time, psychosomatics tended to overemphasize the role of individual differences or dispositions without taking into account the physiological, genetic, and situational factors that interact in predisposing an individual to a particular illness (see Weiner, 1977). Today this perspective has changed, and researchers have become aware of genetically and environmentally determined physiological response patterns that might predispose particular individuals to one disorder rather than another. (This is commonly referred to as the _specificity problem_). For example, studies have shown that individuals differ in the secretion of pepsinogen and that the tendency to secrete this substance, which is related to ulcer susceptibility, may be genetically transmitted (Weiner, 1977).

It should be emphasized that a single predisposing factor may not be enough to result in physical disease. A variety of activating situations are necessary to produce organ dysfunction. This is amply demonstrated by studies of gene-environment interactions in the etiology of high blood pressure and ulcers; it is also shown by the importance of environmental challenge and/or stress in activating Type A behavior. Further investigation of factors leading to the expression of predispositions for specific disorders should contribute to basic knowledge concerning mechanisms of mind-body interaction and provide greater understanding of the physiological processes of disease.

Cancer

Cancer is the second leading cause of death in the United States, accounting for about 20 percent of the overall mortality rate, or nearly 400,000 deaths annually. Despite improvements in the rate of cure, total cancer mortality has risen substantially over the past several decades. This may be attributable, in part, to the growing proportion of older people in the population, since the risk of developing cancer increases with age. Another reason for increased mortality is the dramatic rise in the incidence of lung cancer: Cancer of the lung is the leading cause of death from cancer among men (National Science Foundation, 1980; U.S. Department of Health, Education, and Welfare, 1979a) and may soon have a similar dubious distinction for women.

Cancers are not a single disease; instead, the term is used for more than 100 conditions characterized by unrestrained multiplication of cells and abnormal forms of cell growth (Fraumeni, 1975). One significant attribute of cancers is their ability to spread beyond the site of origin. They may invade neighboring tissue by direct extension or disseminate to more remote locations through the bloodstream or through the lymphatic system, which controls fluid transport between body tissues. It is believed that some or all cancers arise from a single abnormal or transformed cell, triggered in different ways to produce unrestrained multiplication (Levi, 1979). A complementary view is that cancer cells multiply and spread when a breakdown occurs in a portion of the immune system that performs the function of recognizing transformed cells and eliminating them before a detectable tumor can result.

The known risk factors for cancer include a variety of environmental agents, such as tobacco, X-rays and UV-radiation, alcohol, viruses, drugs, asbestos, and many chemicals. Personal attributes, including genetic predispositions, congenital defects, precancerous lesions, and aging, have also been implicated (Fraumeni, 1975). Tobacco is an exogenous substance for which data demonstrate a strong association with cancer: Smoking increases the risk of lung cancer about tenfold, increasing with duration of smoking and number of cigarettes smoked per day (U.S. Department of Health, Education, and Welfare, 1979b). In addition, tobacco use enhances the effects of other carcinogens. For example, exposure to asbestos carries some risk to nonsmokers; however, this is of a low order of magnitude compared with the risks experienced by cigarette smokers. It has been estimated that asbestos workers who smoked cigarettes had eight times the lung cancer risk of smokers without this occupational exposure (U.S. Department of Health, Education, and Welfare, 1979a). This is 92 times the risk of nonsmokers who did not work with asbestos.

While cigarette smoking represents a specific behavior with known pathogenic consequences, three other classes of psychological variables have been suspected as risk factors for cancer: stressful life events, particularly those involving loss (e.g., bereavement); lack of closeness to parents; and inability to express emotions, especially negative ones (Fox, 1978; Schmale, 1981). The research designs used in most of these studies are retrospective. Various methodological flaws tend to render their results suggestive only.

In a prospective study unique in this area, psychosocial data were obtained from 913 male medical students at the Johns Hopkins University long before the clinical appearance of disease (Thomas et al., 1979). A follow-up showed that 20 of the men developed cancer over the next 10 to 15 years. These men had reported a lack of closeness to parents on a family attitude questionnaire taken at the inception of the study. Scores on this measure distinguished the future cancer victims from both the subjects who remained healthy and those who developed high blood pressure or myocardial infarction. However, even this study may be criticized for its statistical methodology. Many variables were measured and only a few yielded significant differences. Moreover, there was a failure to control for known cancer risk factors such as smoking. Replications of these findings must be done before premorbid psychological characteristics can be accepted as risk factors for human cancers.

A related area holding some promise concerns the relationship of psychological factors to cancer growth and progression. This research is a subset of a larger area of study focusing on determinants of successful coping with chronic illness (e.g., Hamburg et al., 1980; Krantz, 1980). Clinicians have often commented on the psychological differences between those cancer patients who do well or survive longer and those who do poorly or succumb rapidly to the disease. Characteristics such as low denial, depression, and anxiety have been related to poor cancer prognosis, and the experience of emotional stress has been observed in patients some months prior to relapse after long remission periods (see Miller and Spratt, 1979).

Research attempting to link psychosocial variables to cancer may be criticized for lacking a theoretical basis. Except for vague reference to the possibility that emotional stress may decrease bodily resistance to malignant growth, little attention has been given to the pathophysiological mechanisms underlying an association between psychosocial variables and the development of cancer. However, experimental research has now provided the groundwork for an investigation of such mechanisms. Advanced techniques for measuring immunological functions show that, rather than existing as an autonomous defense agency, the immune system is integrated with other physiological processes. Moreover, it is subject to the influence of the central nervous system and endocrine responses that accompany psychological stress.

A new interdisciplinary research area, psychoneuroimmunology, examines the interrelationships among central nervous system, endocrine, behavioral, and immunological processes (Ader, 1981). For example, laboratory stressors tend to decrease the responsiveness of the immune system in animals, and stress-responsive hormones, including corticosteroids, can directly and indirectly alter components of the immune response (Ader, 1981; Amkraut and Solomon, 1977). Animal and human studies demonstrate that laboratory and naturalistic stressors can reduce the number of lymphocytes (cells important in the immune process), lower the level of interferon (a substance that may prevent the spread of cancer), and cause damage in immunologically related tissue (Ader, 1981).

Of particular relevance to cancer are other studies demonstrating that stress can inhibit the body's defenses against malignancy. For example, Riley (1975) reported markedly different latencies for mammary tumor development as a function of stress exposure in mice injected with a virus that induces tumors. Those animals housed under conditions of chronic environmental stress (e.g., crowding, noise) developed tumors with a median latency of 358 days, compared with a latency of 566 days in animals housed under protective conditions.

It should be emphasized that the relationship between stress and the immune system is by no means a simple one. Under certain conditions, enhanced immunity and increased resistance to cancers in response to stressors have been reported. Appropriate levels of the hormones released under stressful conditions (e.g., corticosteroids) are essential for normal development and functioning of the immune system (Amkraut and Solomon, 1977). This suggests that the direction of stress effects on the immune system--that is, whether immunocompetence is enhanced or depressed--may depend on the level of stress experienced and resultant changes in hormonal levels.

The exploration of psychosocial influences on immune function constitutes an important area of biomedical research with implications for understanding cancer as well as a variety of infectious diseases (see below). Furthermore, psychoneuroimmunological investigation is needed to isolate the variables that moderate the impact of stress on immunological activity. There is also the possiblity that certain groups of individuals (e.g., the elderly) may be particularly susceptible to psychosocially induced alterations in immune response. This may occur because of documented changes in the immune system (Makidonan and Yunis, 1977) or because of psychosocial changes, such as the decreased financial security and reduced mobility (Eisdorfer and Wilkie, 1977) that accompany aging.

Infectious Diseases

Exposure to contagious microorganisms does not lead invariably to diseases. In fact, only a small percentage of infected persons actually become ill during disease epidemics. There is evidence that psychosocial factors can influence the acquisition, course, and recovery from infectious diseases via at least three mechanisms. These mechanisms parallel the processes linking behavior and physical illness outlined in the introduction.

Health-Impairing Habits and Life-Styles

An individual's behavior can influence exposure to infection and the dose of pathogen. Poor nutrition and/or poor personal hygiene obviously increase susceptibility to illness and delay recovery. In this regard, various behavioral and social factors (e.g., low socioeconomic status) are associated with increased incidence of infectious illness. Individuals in these categories are more likely to be exposed to harmful microorganisms, suffer from known health hazards at home and work, and have poor training in prudent, healthful ways of living. They may also have less access to quality medical care and are less likely to engage in preventive health practices (Institute of Medicine, 1978).

Direct Psychophysiological Effects on Immunity

As noted in the section on cancer, there is evidence that psychosocial factors affect the functioning of the immune system. This leads to increased susceptibility to immunologically mediated diseases and increased expression of the disease among those who are infected. Biobehavioral research has specified certain immunological changes (e.g., reduced production or level of antibodies) that mediate these relationships.

With regard to infections, the immune system can be divided into three functional components: processes involved in transporting the invading microorganisms to the immune system; processes leading to the production of antibodies or immunologically active cells; and processes involving interaction between immunologically active substances and invading microorganisms. Neurohumoral factors can influence each of these immune mechanisms (Amkraut and Solomon, 1977). As with cancer, stress-related influences on susceptibility to infectious diseases depend on a complex of factors. These include the type of stress, the type and number of invading microorganisms, the mode of infection (e.g., air, contact, bloodstream), and the species of animal and its immunological state at the time of inoculation (Ader, 1981; Amkraut and Solomon, 1977).

Psychoneuroimmunological research with animals has demonstrated that behavioral conditioning with noxious stimuli increases susceptibility to viral infections. In one study, neither exposure to a stressor nor inoculations with a virus was alone sufficient to induce disease in adult mice, but the combination of stress and inoculation with a virus elicited symptoms of viral disease (Friedman and Glasgow, 1966). In other studies, rats handled for brief periods of time early in life showed more vigorous antibody response to bacteria than did nonhandled controls (Amkraut and Solomon, 1977). The impact of psychosocial stimulation on the immune response appears to be related to the dose of pathogen administered as well as to the timing of exposure to the stressor (Amkraut and Solomon, 1977).

In humans, much of the evidence linking psychosocial stress to increased susceptiblity to infectious diseases is derived from retrospective clinical studies, although there has been some promising prospective research. For example, Meyer and Haggerty (1962) studied the influence of family crises on factors that might modify susceptibility to streptococcal disease. Over a one-year period, each family member was followed through periodic throat cultures and measures of immunological function. Clinical ratings of chronic stress were positively related to streptococcal illness rate and levels of streptococcal antibodies in the blood. While close contact with infected family members and the season of the year influenced acquisition of a streptococcal organism, several respiratory illnesses were considerably more frequent after family episodes judged to be stressful.

A prospective study of infectious mononucleosis (Kasl et al., 1979) studied a class of military cadets--a population subjected to the rigors of military training and academic pressure. Among subjects susceptible to infectious mononucleosis (i.e., those without Epstein-Barr virus antibodies at matriculation), about one fifth became infected each year with the virus, and one quarter of this group went on to develop the clinical disease. Psychosocial factors that increased the risk of clinical disease among those infected included having a high level of motivation, doing poorly academically, and having "overachieving" fathers.

Behavioral Reactions to Illness

A third process linking behavior to the course of infectious illness involves treatment-seeking behavior and response to treatment. For example, a study of individuals in Maryland who had contracted Asian influenza in 1957-1958 revealed that clinical disease characteristics (e.g., serological response, height of fever, symptom severity) failed to distinguish those who recovered quickly from those who retained symptoms for longer periods of time. However, subjects with delayed recovery scored as more "depression prone" on psychological tests given in advance of the outbreak of illness. This finding was interpreted to indicate that depression-prone individuals exhibit greater concern over illness, which increased and prolonged their physical complaints and reports of illness. A prospective follow-up study measured actual frequency of infection via assays for rises in serum antibody titers. Among those who were infected, depression-prone subjects tended to develop the disease (thus suggesting a possible role of immunological factors), but increased concern over the illness seems the most likely explanation for these findings (Cluff et al., 1966).

TREATMENT AND REHABILITATION OF PHYSICAL ILLNESS

The Prospective Patient and the Medical Encounter

The provision of medical care depends to a very considerable extent on the social, psychological, and cultural processes that lead people to define themselves as requiring care (Mechanic, 1968). Many factors unrelated to the biological severity of illness combine to determine who receives care; persons requiring medical attention do not always seek out medical help and are not always seen by health care providers. These variables must be considered by practitioners and policy planners in the formulation and delivery of health care service.

Recognition of symptoms and the resultant use of health care services are influenced by situational factors, such as life difficulties and psychological stress (Mechanic, 1968). Also important are learned patterns of behavior, such as social roles and cultural norms; for example, women are more likely than men to visit health care professionals. Social class and cultural background influence patients' evaluations of symptoms and doctors' responses to patients' complaints. Age is another factor that determines reactions to symptoms and the use of medical facilities; for example, the elderly take aches and pains for granted and place little faith in medical science, even though the frequency of use of medical facilities and concern with health increase with age (Riley and Foner, 1968).

Psychological and Physiological Aspects of Pain and Illness

Much progress has been made in identifying sociopsychological correlates of pain and psychophysiological pain mechanisms and in developing research-based techniques for pain control. Pain is more than a sensory experience. It is not a necessary consequence of injury or tissue damage. Definitions that imply that pain can be stopped simply by interrupting neural pathways are not adequate to account for clinically related phenomena. For example, surgical interventions indicate a rather disappointing record of success (Weisenberg, 1977). People without known organic pathology suffer pain (Fordyce, 1976), and even when an organic basis for pain is established, psychological factors continue to affect the experience of pain.

A range of cultural, sociopsychological, and situational factors influence pain perception and tolerance (see Weisenberg, 1977). Different cultural groups have different views of appropriate pain reactions, including the circumstances under which it is permissible to cry or ask for help. Moreover, the influences of the social context and the meaning of the pain experience produce differences between clinical and experimentally induced pain. Pain in clinical situations involves anxiety associated with the disease process and fear of death (Beecher, 1959; Weisenberg, 1977), whereas experimentally induced pain does not.

The Placebo Effect

A pervasive phenomenon in the pain literature is the placebo effect, that is, the reduction of pain or the removal of symptoms via medication or therapeutic treatment that has no identifiable active component (Shapiro, 1971). It has been estimated that placebo medication (e.g., pharmacologically inert substances) and other unspecific treatment factors reduce pain successfully in about 35-40 percent of patients (Beecher, 1959). Although the placebo response in medicine has been widely recognized, until recently it had been regarded as a nuisance variable. In pharmacological research the routine inclusion of a control group receiving an inert medication has been considered an essential methodological control, particularly in evaluating psychoactive drugs. Mechanisms of the placebo response as a component of various therapeutic interventions have now come under study in their own right. Contrary to the popular belief that placebo effects are confined to psychological changes, there are data showing that placebos can produce a variety of changes on a physiological level--for example, significant blood pressure reductions among persons with hypertension (Shapiro, 1982).

Psychosocial variables that enhance the effectiveness of placebos have been identified, thereby shedding light on the mechanisms of placebo action in pain relief. Person-centered approaches aimed at identifying patients who are responsive to placebos have not proven valuable. There is evidence, however, that situational factors that influence a patient's motivational and attentional processes as well as a host of variables relating to doctor-patient interaction (e.g., expectations of relief, the patient's confidence in the physician and the procedures), can heighten placebo effects (Shapiro, 1971). Stress and anxiety reduction also seem to be an important facet of placebo effectiveness (Beecher, 1959). It is likely that placebo-related factors, when fully understood, will provide a powerful tool in clinical practice.

Psychophysiological Models

The influential gate control theory of pain was proposed 15 years ago (Melzack and Wall, 1965) in order to integrate physiological and psychological factors in pain perception. This theory proposes that noxious stimuli activate selective central nervous system processes, which act to exert control over incoming messages. Influenced by this trigger mechanism, cells at each level of the spinal cord act as a gate control system, increasing or decreasing their receptivity to incoming pain signals traveling along the nerves. This system makes it possible for the higher mental processes, which underlie attention, emotion, and memories of prior experience, to alter the transmission of pain signals (Weisenberg, 1977).

The posited physiological and anatomical bases for the gate control theory have been subject to considerable criticism (e.g., Liebeskind and Paul, 1977; Nathan, 1976). However, a wide assortment of clinical and experimental findings have been interpreted as supporting the theory--or at least certain aspects of it (Liebeskind and Paul, 1977)--and the theory has led to the development of a technique for artificially stimulating the nervous system to relieve pain. Although subsequent work on endorphins calls into question some key details of the gate control theory, the original theory and subsequent modifications have been influential in highlighting the importance of motivational and cognitive factors in the experience of pain.

The recent discovery of endogenous, opiate-binding receptors and substances in the brain (endorphins) that bind with these receptors has led to a new interest in central nervous system mechanisms of pain control. It has been demonstrated that a pain-suppression system exists in the brain that can be activated by psychophysiological procedures such as electrical stimulation (Liebeskind and Paul, 1977) and by environmental and psychological manipulations such as exposure to stress. Recent research (Levine et al., 1978; Mayer et al., 1976) suggests that mechanisms of action of heretofore poorly understood phenomena of pain relief (e.g., acupuncture, placebo response) may involve this endogenous system.

The physiological involvement of endorphins is investigated by administering naloxone, a drug that specifically blocks the action of opiates. If an endogenous physiological process is blocked by naloxone, endorphins can be inferred to play a role in this process. Levine et al. (1978) investigated the possible role of endorphin activity in placebo pain relief with patients who had just undergone painful dental surgery. Naloxone was administered to half the patients under randomized, double-blind conditions; the other half received a placebo. Pain ratings were taken, and those patients previously given a placebo were further subdivided into two groups: those whose pain was reduced or unchanged (placebo responders), and those whose pain increased (nonresponders). Patients were then randomly given a second drug, again either naloxone or a placebo. Results indicated that after the first drug administration, subjects given naloxone reported more pain than those given a placebo. Naloxone given as a second drug produced no additional increase in the pain levels of placebo nonresponders but did increase the pain levels of placebo responders. These data are consistent with the hypothesis that endorphin release mediates placebo pain relief for dental postoperative pain. This study is only an initial investigation of the possible role of endorphin activity in psychological pain phenomena. Future research on endorphins promises to have a profound impact not only on basic science and clinical understanding of pain, but also on a wide range of biobehavioral phenomena, including addictive behaviors and mental illness.

Compliance with Medical Regimens

In recent years there has been a growing awareness that the failure of patients to adhere to prescribed medical regimens is probably the single greatest problem in bringing effective medical care to the individual patient. This problem also contributes in a major way to the economic and social costs of illness (Cohen, 1979; Sackett and Haynes, 1976). Although adherence varies, it is not uncommon to find compliance rates as low as 50 percent in many situations. It was estimated recently that only one third of patients comply, that one third are noncompliant because they adhere to a misunderstood regimen, and that one third are knowingly noncompliant (Cohen, 1979). The medication compliance problem may be most pronounced for chronic illnesses, such as high blood pressure, for which effective therapy requires regular, long-term taking of medications that may produce unpleasant side effects. (Sackett and Haynes, 1976). However, noncompliance is a problem in areas of the treatment process other than adherence to prescribed regimens. Substantial numbers of patients who do not have painful symptoms fail to come for scheduled appointments, and the problem of inducing and maintaining change in unhealthful habits (such as diet and smoking) is particularly formidable.

A good deal of attention has been given to isolating factors that influence or predict compliance (Becker, 1979; Sackett and Haynes, 1976). Surprisingly, common demographic variables such as age, sex, and marital and socioeconomic status have little independent influence. The crux of the problem is often poor doctor-patient communication, rather than the patient's behavior alone. Two sets of variables deriving from the physician-patient encounter--satisfaction with care and comprehension of treatment regimen--appear to affect compliance.

Aspects of the doctor-patient relationship determine satisfaction, and satisfaction determines the degree to which medical advice is accepted. For example, in a study of a pediatric setting, Korsch and Negrete (1972) found that a major source of mothers' dissatisfaction was the failure of physicians to answer questions and provide clear explanations of illness. More than 80 percent of those who thought the physician had been understanding were satisfied, compared with only one third of those who did not feel that the doctor tried to understand their problems. If mothers were dissatisfied with the doctor or the content of the consultation, they were less likely to comply with the advice.

A second aspect of the compliance problem is the patient's ability to comprehend and recall details of the treatment regimen. Much of the failure to follow doctor's orders stems from genuine problems in understanding and remembering what is said (Ley and Spelman, 1967). Often the material presented by the doctor is too difficult to understand, the treatment regimen itself is overly complicated, or patients hold misconceptions about illness or human physiology that lead to confusion.

The causes and consequences of comprehension problems are illustrated by a study in which physician-patient interactions were observed for a sample from a lower-class clinic (Svarstad, 1976). Reviews of medical records and pharmacy files, follow-up interviews, and validation of patients' reported behavior via pill counts were made. The data revealed that patients did not always leave the clinic with an accurate perception of what physicians expected them to do. Frequently physicians were not explicit in discussing their expectations with the patients. When physicians did make efforts to motivate compliance by being friendly, appealing to reason, or checking on previous compliance, many more patients conformed to the prescribed regimen.

The crucial challenge for medical compliance, as with the modification of other health-impairing behaviors, is to maintain people on prescribed regimens for sustained periods. This problem is illustrated by the remarkably similar relapse rates among subjects treated in programs aimed at weight reduction, smoking cessation, and reduction of alcohol consumption (Hunt et al., 1979). Two thirds of such patients abandon the regimen and backslide by the end of three months, and only about one quarter of the individuals maintain changed behavior at the end of a one-year period.

One technique used to help maintain long-term adherence to treatment regimens is a focus on the immediate rewards and consequences of compliance or noncompliance. A range of behavior modification procedures, based on principles of operant conditioning, have proven most effective (Pomerleau and Brady, 1979). Interventions to increase adherence must also recognize those characteristics of the social interactions of physicians and patients that foster noncompliance.

Most of the research on medical compliance has been designed to solve practitioners' everyday clinical problems, rather than to develop a comprehensive theory that may apply across a broad range of medical situations, illnesses, and behaviors. However, one conceptual approach that has received some support in explaining medically related behaviors (including compliance) is the Health Belief Model. This model centers on the patient's views about the appropriate paths of action in the presence of health disturbances, perceptions of barriers to action, and subjective interpretations of symptoms (Becker, 1979). Still more effective approaches are needed that encompass the physician-patient communication process and suggest ways of making the rewards of long-term medical compliance more salient to patients.

Smoking: An Example of Health-Impairing Behavior

The cigarette smoking habit has been described as the single most preventable cause of death in the United States (U.S. Department of Health, Education, and Welfare, 1979a). Despite the fact that knowledge of the health risks of smoking has reduced the percentage of adults who regularly use cigarettes, there are still over 50 million smokers in the United States today. Moreover, in recent years there has been an alarming increase in the proportion of teenagers (particularly girls) who are taking up the smoking habit. With regard to the modification of smoking behavior, the problem is not that the public is unaware of the negative health consequences, but that the great majority of smokers are unable to quit or stay off cigarettes for prolonged periods (Bernstein and Glasgow, 1979; Leventhal and Cleary, 1980).

Cigarette smoking is a behavior whose initiation, maintenance, and cessation are determined by a mixture of social, psychological, and physiological factors. Many of the problems faced in attempts to prevent and modify the smoking habit are associated with the correction of other health-impairing habits, life-styles, and dependencies (e.g., poor diet, alcoholism, lack of exercise). We give considerable attention to the smoking problem to illustrate behavioral science approaches to these health-impairing behaviors.

Initiation and Prevention of Smoking

Cigarette smoking can be viewed as the product of a multi-stage process that begins with initial experimentation and leads to the acquisition of a habit and/or addiction (Pomerleau, 1979). Data suggest that even limited adolescent experimentation may lead to habitual smoking (Leventhal and Cleary, 1980). Psychosocial factors related to the initiation of smoking include social pressure from peers, imitation of adult behavior, adolescent rebellion and antisocial tendencies, and personality factors such as extraversion--a biologically based dimension related to arousal or stimulation-seeking. A social learning explanation of smoking initiation (Bandura, 1977) assumes that the habit is acquired through imitation and social reinforcement, typically under the influence of peer pressure, media stereotypes, etc.

The inhalation of smoke is initially somewhat aversive, but after sufficient practice, pharmacological habituation (or tolerance) occurs, and the behavior produces enough satisfaction or reward in its own right to maintain the habit (Pomerleau, 1979). The immediate social and biological rewards of smoking (and the delay of negative health consequences) may account for many of the difficulties in modifying the habit once it becomes established (Jarvik, 1979).

Early efforts to prevent smoking assumed that this could be accomplished by teaching young people the health consequences of the habit, but the results were largely disappointing. However, several projects have obtained encouraging results by employing sociopsychological techniques of communication and attitude change to deter smoking in adolescents. A pioneering effort in the area, the Houston Project, is a three-year longitudinal study (Evans et al., 1981). This project created persuasive films and posters to teach young teens (grades 7-9) about peer and media pressures to smoke and about effective techniques for resisting pressures. Other films demonstrated the immediate physiological consequences of smoking (e.g., carbon monoxide on the breath). Hundreds of students in matched experimental and control groups were compared for cigarette smoking rates at the start of the project and during the three-year follow-up. The results indicated a significant impact of the films and posters: Experimental subjects smoked less frequently and expressed less intention to smoke compared with a control group receiving no intervention (Evans et al., 1981).

Maintenance of the Smoking Habit

Once smoking is established, both psychological and biological factors contribute to its persistence and resistance to change. Learning mechanisms, possibly in conjunction with the physiological satisfactions derived from smoking, play a considerable role in maintaining the habit

(Hunt and Matarazzo, 1970). The use of cigarettes becomes part of a chain of behaviors: taking out the package, lighting the cigarette, getting tobacco smoke, etc. As a result, these stimuli associated with smoking come to elicit pleasurable responses by themselves. In addition, the avoidance of unpleasant withdrawal effects (e.g., craving) becomes rewarding, thus helping to maintain the habit (Russell, 1979). These observations receive more systematic support from animal research, which demonstrates that drug responses (e.g., morphine tolerance) can become conditioned reactions and that withdrawal symptoms can be conditioned to external cues (Siegel, 1979).

Learning or conditioning mechanisms alone are not sufficient to explain the maintenance of smoking, since many smokers increase intake to regulate or achieve a particular level of nicotine in their system (Schachter et al., 1977). Biological factors figure prominently in the maintenance of the habit, and nicotine is the chemical in tobacco that is probably most responsible for these effects (Jarvik, 1979). However, the question of whether cigarette smoking can be considered an addiction comparable to heroin or alcohol addiction remains a subject of scientific debate (Russell, 1979).

Tobacco has the capacity to elicit many of the defining characteristics of an addictive process, and recent biobehavioral research has focused on the complex interplay of psychological and pharmacological processes leading to smoking behavior. For example, a nicotine regulation hypothesis asserts that heavy smokers adjust their smoking rate to keep nicotine at a roughly constant level, and that the rate of smoking depends on the rate of nicotine excretion and breakdown by the body. The rate of nicotine excretion depends in part on the acid-base balance (pH) of the urine, which in turn can be altered by psychological stress or anxiety. Thus it is argued that the links between psychological processes, the craving for cigarettes, and increased smoking are mediated by a physiological addiction mechanism involving the pH of urine (Schachter et al., 1977).

A series of programmatic studies (Schachter et al., 1977) provide support for this hypothesis. First, a sample of longtime heavy smokers consistently smoked more low- than high-nicotine cigarettes, thus showing that smokers "regulate" nicotine intake. Next, it was shown that when urinary pH was manipulated by the administration of acidifying or alkalizing agents, smokers smoked more when urine was acidified. A third set of studies examined the urinary pH-excretion mechanism as a mediator of psychological determinants of the smoking rate. Urinary pH was found to covary with naturalistic situations (e.g., party-going, examinations) associated with heavier smoking. Schachter's point is that psychological variables (such as stress) influence the smoking rate only because of their effects on the urinary pH mechanism. Support for this was provided by a laboratory experiment that independently manipulated stress and urinary pH; results indicated that smoking covaried with pH, rather than stress.

There remain several exceptions to the nicotine regulation model. There are different types of smokers, some of whom do not appear to smoke for nicotine content (Schachter et al., 1977). In addition the Schachter model suggests that the lowered nicotine content of cigarettes will increase the number of cigarettes smoked, and data in this regard are at best contradictory (Garfinkel, 1979). Nevertheless, the nicotine regulation hypothesis has identified a biobehavioral mechanism for cigarette addiction.

Withdrawal

Because of the addictive component of smoking, it is crucial to understand the withdrawal process in order to develop effective intervention strategies to modify the habit. However, most research efforts have concentrated on the effects of cigarette smoking rather than on the effects of cessation--irritability, sleep disturbances, inability to concentrate, and weight gain (Schachter, 1978).

Recent research by Grunberg (in press) suggests that weight gain accompanying withdrawal from nicotine may result from increased preferences for sweet foods. Smokers allowed to smoke ate fewer sweets than did nonsmokers or deprived smokers. The three groups did not differ in consumption of nonsweet foods. A parallel study with animals showed that nicotine administration retarded normal body weight increase in rats. Cessation of nicotine was accompanied by marked increases in body weight and concomitant increases in consumption of sweet foods. Moreoever, these effects could not be explained by changes in total food consumption or activity level.

Ability to cope with the withdrawal syndrome is crucial to the maintenance of smoking cessation. Therefore, further investigation of mechanisms responsible for symptoms accompanying withdrawal may suggest techniques for controlling the high recidivism rate among those who attempt to quit smoking.

Modification of Smoking Behavior

Most of the effort in modifying smoking behavior has been directed toward the development of smoking cessation strategies. Many of the earlier intervention studies suffered from problems of experimental design (e.g., lack of adequate control groups) and difficulties in measuring smoking cessation objectively. There was also a high early dropout rate, sometimes reaching 50 percent of subjects included in the initial sample, which may spuriously inflate the initial success rate (see Leventhal and Cleary, 1980; Pomerleau, 1979). More recent work has enabled some systematic evaluation of the long-term efficacy of smoking cessation programs and techniques.

Therapy Approaches Therapy approaches include individual and medical counseling, hypnosis, group therapy, and behavioral therapies derived from the learning theories of experimental psychology.

Research indicates that most therapy techniques are effective in promoting short-term cessation of smoking but usually fail to keep more than 50 percent of ex-smokers off cigarettes for long periods of time (Bernstein and Glasgow, 1979). Systematic study of the important area of maintenance of nonsmoking is just beginning, and strategies such as long-term group support and behavioral techniques for coping with anticipated withdrawal symptoms are promising. A number of the public health studies described below incorporate components of behavioral therapy approaches. Successful long-term smoking cessation results obtained by these studies are due in part to therapy techniques.

Public Health Approaches It appears that significant reductions in adult smoking, especially among middle-aged men and certain professional groups, can be attributed to information and educational campaigns initiated after the first Surgeon General's report on smoking in 1964 (Pomerleau, 1979). In recent years several large-scale media-based projects have been undertaken in the United States and Europe to change attitudes and behavior related to smoking.

The Stanford Heart Disease Prevention Project was designed to reduce a broad range of risk factors, including smoking (Farquhar et al., 1977; Meyer et al., 1980). Three communities were studied: one served as a control; a second was exposed to a mass media campaign on heart disease risk factors, including smoking; and a third received the mass media campaign and face-to-face behavioral therapy for selected high-risk persons. The media campaign alone produced some reductions in smoking at long-term follow-up. More substantial reduction occurred when the media campaign was supplemented by face-to-face therapeutic instruction. These findings are encouraging but must be evaluated cautiously because of several methodological problems inherent in studies of risk factors of this sort--namely, the high dropout rate and/or other difficulties encountered when subjects do not adhere to randomly assigned interventions involving life-style changes (Kasl, 1980).

Another ambitious study in Finland (Puska et al., 1978) introduced a nationwide multiple-component program, including televised counseling sessions. These were designed to prevent relapse by educating participants in behavioral techniques for coping with anticipated relapse problems (e.g., stress, weight gain). About 40,000 adult smokers participated in the study; as a result of the program a small but significant percentage achieved sustained abstinence from smoking at six-month and one-year follow-ups (McAlister et al., 1980).

The Stanford and Finnish programs represent impressive and perhaps cost-effective efforts to induce large numbers of people to abandon the cigarette habit on a long-term basis. These and related studies, however, have not produced unequivocally successful outcomes (Leventhal and Cleary, 1980; Kasl, 1980).

Nonetheless, they do suggest that meaningful changes in smoking behavior via public health approaches are possible, but only when the risks of smoking are made immediate and salient, and both skills and support to change smoking behavior are provided (Pomerleau, 1979). The more important question of whether the reduction of risk factors will lower morbidity and mortality, particularly from cardiovascular diseases, is being studied directly by large-scale intervention trials now under way. These projects are discussed in the last section of this paper.

Other Behavior Therapies in Health Care

The increasing importance of behavioral and social sciences in medicine is due in part to the development of effective procedures for changing illness-related behaviors (see the paper by Wilson in this volume). Several of these behavior modification techniques, which were designed and evaluated for the prevention, management, and treatment of physical disease (Pomerleau and Brady, 1979), were alluded to in the discussions of medical compliance and cigarette smoking. Other health care applications are in the areas of pain control, childhood disorders, adult psychosomatic disorders, rehabilitation of the disabled and physically ill, and geriatric problems (Melamed and Siegel, 1980). Four representative behavioral techniques are described below: the operant control of pain, cognitive-behavioral interventions, biofeedback, and relaxation training.

Pain reactions can persist long after the original physiological sensation and tissue damage have been remediated. Fordyce (1976) and others have developed a successful technique for treatment of chronic pain through the application of operant conditioning procedures. Many pain-related behaviors become established and maintained by the particular rewards they provide for the patient--for example, attention from family and staying home from work, as well as pain relief. These rewards are therefore manipulated so that the value of undesirable pain behaviors is reduced or removed.

Cognitive-behavioral interventions are techniques designed to reduce pain and the aversiveness of medical procedures by diminishing the perceived threat and the psychological stress associated with the procedures (Turk and Genest, 1979). Since the physiological and/or behavioral components of the stress response (e.g., excessive sympathetic nervous system activity, lowered motivation or ability to comply with medical regimens) may also interfere with the recover process (see Krantz, 1980), there is some indication that stress reduction procedures can speed recovery. Some of these procedures (e.g., psychological preparation of children for hospitalization) are being applied routinely (Melamed and Siegel, 1980).

Until recently it was believed that the responses of the autonomic nervous system were involuntary and that an individual could exert little or no control over these processes. However, visceral responses such as heart rate, blood pressure, and skin temperature can be controlled voluntarily when feedback is provided to the individual for altering these responses (see Miller, 1969). Biofeedback training teaches the individual to monitor physiological responses through the use of electronic instruments. When a subject alters a physiological state (e.g., heart rate, muscle tension, electrical activity of the brain), he or she is provided with auditory, visual, or other feedback indicating that the correct response has been made. The feedback is effective in teaching subjects to control the physiological response because it tells them that their motivated attempts to alter the response are effective. Thus, the feedback serves as a reinforcer (reward), which leads to learned control of the physiological response (Miller, 1969).

Recent research has explored the clinical utility of biofeedback techniques for such disorders as high blood pressure, migraine headache, seizure disorders, sexual dysfunctions, and muscular paralysis (Gatchel and Price, 1979; Ray et al., 1979). A particularly effective clinical application has been in the treatment of neuromuscular disorders. For example, through the use of feedback for the activity of muscle units (cells), paralyzed or damaged muscles may once again come under voluntary control. Dysfunctions such as cerebral palsy, muscular spasms, and various paralyses have been successfully treated by biofeedback. Previously these dysfunctions were often unresponsive to traditional physiotherapies and medical or surgical treatment (Ray et al., 1979).

The initial enthusiasm for biofeedback probably exaggerated its therapeutic effectiveness. Further research has revealed limitations in the use of this technique. For example, it must still be established that training in laboratories or clinics generalizes to real-life settings, and more research is needed to evaluate the relative effectiveness of biofeedback versus other therapeutic techniques.

Relaxation therapies are procedures designed to elicit physical and emotional calmness in order to decrease autonomic nervous system arousal, muscular tension, and other physiological correlates of psychic trauma. The most common technique is deep muscle relaxation (Jacobson, 1938), which involves supervised practice in the systematic relaxation of major skeletal muscle groups. Relaxation therapy is effective in the treatment of a variety of psychophysiological disorders, including high blood pressure, migraine headache, and chronic pain syndromes (Melamed and Siegel, 1980).

PROSPECTS FOR FUTURE RESEARCH

An important priority for research in the next few years is the integration of behavioral and biomedical knowledge in a way that elucidates the mechanisms underlying the interplay among behavior, physiological processes, and somatic dysfunctions. Accordingly, the key issues for biobehavioral inquiry include further study of features of the behavioral context and of the individual (e.g., coping styles, biological predispositions, availability of social supports), which may determine the outcome of exposure to stressful events. Also suggested are further studies of psychophysiological mechanisms that mediate linkages between behavior and disease, particularly those involving neuroendocrine and immune responses. Other priorities are the development and evaluation of techniques to produce sustained changes in behavioral risk factors. This includes research on mechanisms of smoking addiction and withdrawal and the prevention of health-impairing habits. The important area of medical compliance requires more theoretically based research taking into account doctor-patient communication and the cognitive and motivational factors that sustain adherence to treatment regimens.

Biobehavioral Paradigm for Research on the Etiology and Pathogenesis of Physical Disease

Psychosocial Stress

As noted earlier, stress has been implicated as a central factor in the etiology of cardiovascular illness and also may play a role in the development of peptic ulcer, cancers, and infectious diseases. The association of psychological stress with somatic disorders underscores the importance of research aimed at (1) understanding when and under what conditions stress becomes translated into physical diseases; (2) specifying the physiological and neuroendocrine pathways through which reactions to stress potentiate illness; and (3) identifying factors that predispose individuals to one stress-related disorder rather than another (see Graham, 1972).

Conditions for the Stress-Disease Relationship Stress is probably an inevitable aspect of modern living; the struggles, conflicts, and frustrations that threaten individual well-being seem to be an inherent quality of the human condition. Yet, stress-related diseases are far from universal. It would appear that personal attributes and contextual variables exert an important influence in modifying the outcomes of psychological stress (F. Cohen et al., 1980). Among the more promising of these modifiers are sociocultural resources, including direct help or emotional support from other people and the health care system. Also important are particular psychological characteristics of stressors, such as whether they are predictable or within the individual's ability to control. In addition, biological predispositions and acquired factors such as styles of coping may mediate the impact of stressors, thereby influencing the likelihood of a disease outcome (F. Cohen et al., 1980). Thus, univariate studies linking indices of stress end points must be supplanted by multivariate research in which the stress-moderating effects of biological, psychological, and sociocultural variables are taken into account.

For research to be cost-effective, psychosocial variables must, when appropriate, be examined in conjunction with ongoing biomedical studies. For example, the National Institutes of Health are currently funding a number of Specialized Centers of Research (SCOR), which are concerned with various aspects of cardiovascular disorders in children and adults (U.S. Department of Health, Education, and Welfare, 1978). Given the importance of behavioral variables in these disorders, valuable scientific data at a relatively low cost would be provided by incorporating behavioral components in these studies. For example, a psychosocial component was included in the Framingham Heart Study cohort examined in the late 1960s. Results of this study, reviewed above, provided useful information supporting the importance of behavioral and social variables in the etiology of coronary heart disease (see Haynes et al., 1980).

Psychophysiological Mechanisms Research on the psychophysiological pathways linking stress to disease will require continued technological improvements to facilitate the measurement and identification of neuroendocrine, central nervous system, and related processes. Animal models will play an important role in such research (see Ader, 1976; Campbell and Henry, 1982). The study of pathophysiological mechanisms often relies on procedures that cannot be used with human subjects. These procedures include the use of surgical interventions and electrical stimulation as a means of identifying sites that regulate bodily reactions to stressful events. Drugs that selectively stimulate (or block) the activity of suspected mediating structures, such as the receptor sites of the sympathetic nervous system, provide a direct means of assessing the impact of stress-related physiological processes on target organs whose dysfunction is suspected to be of psychogenic origin (e.g., Obrist, 1981).

Human research models will remain indispensable, however, especially in the study of cognitive and perceptual variables that initiate and regulate physiological reactions to stressful stimuli. Experimental research is essential and justifiable for making progress in this area in which no demonstrable damage to subjects can be discerned. Where ethical and practical concerns limit the applicability of laboratory methodologies in studying a problem area, it is frequently possible to conduct studies of populations who are exposed to the variable of interest under natural conditions.

Recent developments in psychophysiological measurement make it possible to measure the influence of behavioral variables on physiological processes in natural settings, such as home or workplace. These techniques have opened new frontiers in biobehavioral research. For example, a recent study successfully used a portable electronic device to provide blood pressure biofeedback aimed at preventing fainting, thereby aiding in the rehabilitation of paralyzed patients (Miller, 1979). Another study (Dimsdale and Moss, 1980) used a portable blood withdrawal pump to monitor plasma hormone levels during periods of emotional stress and exercise.

A focus on mechanisms linking behavior and health is required in order to translate historical and epidemiological descriptors, such as age, personality, genetics, or nutritional history, into psychophysiological processes that can be modified or altered (Schwartz et al., 1979). To influence medical practice, behavioral and social science research must identify modifiable variables involved not only in the etiology of disease, but also in the progression of illness after symptoms have appeared (Stachnik, 1980).

The Specificity Problem The study of factors that selectively predispose individuals to particular disorders must incorporate examination of both traditional risk factors (including genetic predispositions) and acquired behaviors (such as coping styles) as well as features of the situation (e.g., exposure to particular types of stressors).

Both animal and human research models need to be used in exploring issues of selective susceptibility to disease. The controlled breeding of infrahuman species can facilitate the study of the genetic and behavioral interplay. For example, several strains of rats susceptible to stress and salt-induced hypertension have been produced through selective breeding (Campbell and Henry, 1981). Similarly, genetic strain appears to influence susceptibility to gastric lesions caused by experimental immobilization (Weiner, 1977).

Human research should also contribute to the study of biobehavioral factors that make for vulnerability to specific physical diseases. Recall that subjects with a family history of hypertension exhibit enhanced blood pressure response while working at a demanding task (Obrist, 1981). A similar research strategy could be used to study physiological changes in subjects with a family history of other disorders. For example, stress-induced changes in serum glucose levels might be studied in individuals with diabetic parents. This type of research should be supplemented by studies of twins and prospective research employing longitudinal designs.

Type A Behavior Pattern

One illustration of a developing area of mechanism-oriented biobehavioral research is the study of the Type A "coronary-prone" behavior pattern. Having demonstrated its association with coronary disease, research now is addressing issues similar to those discussed in the section on psychological stress: (1) isolation of aspects of the behavior pattern that confer enhanced risk; (2) identification of the psychological mechanisms that produce and sustain coronary-prone behavior; and (3) specification of the physiological processes that account for the enhanced risk of individuals displaying coronary-prone behavior (Glass, 1981). Subsequent studies (probably with animal models) might be undertaken to elucidate cause and effect. That is, do animals bred or trained to exhibit Type A behavioral characteristics show elevated physiological reactivity, or are the behavioral responses caused by physiological reactivity? Indeed, both behavioral and physiological reactions may be consequences of a third variable located elsewhere in the nervous system.

Psychoneuroimmunology

The emerging field of psychoneuroimmunology also holds great promise (see Ader, 1981). Exploration of basic mechanisms of immune changes produced by psychological stimuli will continue to be an active area of research. In addition to controlled laboratory experimentation with animals, there is a need to determine if reliable, replicable, and clinically meaningful alterations in immune function in humans are associated with psychosocial variables (e.g., certain life stressors, coping styles, or both of these acting together). Other research priorities for this field include the study of correlated changes in neuroendocrine and immune functions across the life-span (developmental immunology); studies of possible learning and conditioning effects on the immune system; and prospective studies relating behavior to processes of immunologically mediated diseases (Ader, 1981).

Methodological Issues

The complexities involved in integrating behavioral and biomedical knowledge will require multifaceted research strategies. What is needed is a continual interplay between laboratory and field methodologies. This interplay may take several forms. For example, an effect can be established as reliable with controlled laboratory experimentation, in which causal links can be inferred. The generality of the relationship can then be established in subsequent research in natural settings such as the home or workplace (see S. Cohen et al., 1980). Similarly, by first conducting field studies, it is possible to isolate important dimensions of a particular research area. At that point laboratory studies may be useful to rule out alternative explanations often inherent in naturalistic research. A vivid example of this methodological interplay is provided by data on biobehavioral factors in the etiology of high blood pressure. Naturalistic and clinical evidence suggested that psychosocial stress plays a role in this disorder. Accordingly, laboratory studies were undertaken to isolate the psychophysiological mechanisms involved in behavioral responses to environmental stressors. Further naturalistic work (e.g., Rose et al., 1978) extended the laboratory findings by demonstrating that exaggerated blood pressure responses to high work loads were predictive of sustained hypertension (Herd, 1978).

Risk Factor Modification and Prevention

Associations between major chronic diseases and seemingly modifiable behavioral factors have spurred interest in relating behavioral knowledge to the promotion of health and the prevention of disease (Breslow, 1978; Matarazzo, 1980). The present body of research in this area constitutes only a promising beginning, and it is wise to be cautious about making unequivocal claims of success based on existing evidence. However, this emerging research area does raise important challenges and questions.

Maintaining Abstinence

While there are encouraging indications that established patterns of behavior can be changed in the short term, a major difficulty has been maintaining these changes in substantial numbers of individuals over sustained periods of time (Bernstein and Glasgow, 1979; Hunt et al., 1979). There is also a high early dropout rate in various treatment programs (see Leventhal and Cleary, 1980). Work in these areas by behavioral scientists will intensify in the next five years and must focus on understanding the factors that initiate and maintain health-impairing habits, not just on techniques to modify and prevent them.

For some habits, such as smoking or drug abuse, biological factors are intimately involved at all stages of the problem. Considerable attention must be given to the psychobiological and psychosocial aspects of the processes of withdrawal and behavior change themselves (Leventhal and Cleary, 1980). Smoking, diet, and exercise habits and health-endangering practices such as failure to use seat belts, alcohol abuse, and poor hygiene must also be studied as sociocultural phenomena. Decisions to engage in or modify health-impairing habits and the incorporation of changed behaviors as part of an overall life-style all occur in a social context (Syme, 1978).

Antecedents of Habits and Risk Factors

Habits and life-styles develop in the context of family and society; hence, more research is needed on the socialization of health-related habits. Such longitudinal and cross-cultural research is expensive, but it may be conducted in a cost-effective manner in conjunction with ongoing longitudinal studies of the development of disease risk factors in children. For example, a number of projects are being carried out among populations of school-age children (e.g., the Bogalusa Heart Study, Voors et al., 1976) to track the distribution and time course of risk factors of heart disease such as blood pressure and serum lipids. Behavioral and social variables, including family health values and habits, could be incorporated into such projects. A behavioral interface with biomedical research would also provide an excellent opportunity to examine the processes involved in the socialization of health life-styles.

Prevention

Primary prevention of health-impairing habits (i.e., before disease develops) and the promotion of healthful life-styles for people of all ages are cost-effective approaches to health, for in the long term the potential costs, in lives and dollars, of treating disease are likely to outweigh the costs of preventing unhealthful habits. Social learning approaches to smoking prevention have yielded promising results in the Houston school-based intervention (Evans et al., 1981). Further work with children and adolescents might expose other habits to social learning interventions. More systematic research with adults is also needed. The workplace has proven to be a promising setting for such efforts. People spend considerable time at work, and many employers sponsor such programs because of the benefits that accrue from having healthier employees.

The terms secondary prevention and tertiary prevention refer, respectively, to interventions taken to arrest the progress of illness already in early asymptomatic stages, and interventions to stop the progression of a clinically manifest disease (Institute of Medicine, 1978). Secondary and tertiary prevention activities involving behavioral factors may be more feasible than primary prevention, given the current state of knowledge. Advantages of such interventions are that target groups can be easily recognized and are motivated to change their behavior (see Institute of Medicine, 1978).

Modification of Type A Behavior

Various therapeutic approaches have been proposed for modifying Type A behavior (Roskies, 1980). Behavioral techniques such as relaxation training have been proposed as a way of reducing stress-related bodily responses elicited in Type A individuals. Other strategies for modifying Pattern A have been designed to induce behavioral change. One such technique involves having the subject imagine situations that normally elicit Type A behaviors and covertly rehearse alternative, Type B responses. Group therapy procedures also have been used in some studies with patients following myocardial infarction. Efforts to evaluate the effectiveness of these procedures have yielded encouraging results, but care must be exercised in drawing definitive conclusions. Preliminary evidence suggests a reduction in cardiovascular complications and in Type A behaviors (Friedman, 1979; Roskies, 1980).

Systematic research aimed at assessing modification procedures for Type A behavior is certainly one of the priorities in this area. However, large-scale trials may be premature at this time. A more pressing priority is to delineate the particular features of the behavior pattern that are risk-enhancing as well as the psychological factors that give rise to and sustain Type A behavior.

Determining the Impact of Behavior Change on Morbidity and Mortality

The presumably causal associations between behavioral factors and chronic diseases imply that effective modification of habits and behavior patterns will reduce the incidence of and mortality from these disorders. The assumption is complex and requires further evidence before it can be accepted. In the case of cigarette smoking, epidemiological data reveal that former cigarette smokers experience declining overall mortality rates as the years of discontinuance of the habit increase (U.S. Department of Health, Education, and Welfare, 1964). Data of morbidity are more complex and indicate that the benefits of being an ex-smoker are not as high as the benefits of never having smoked. Similarly, the data on the effects of reduced blood lipids on coronary heart disease are not conclusive (Kasl, 1980). Indeed, they suggest that factors such as the age at which reductions occur and the underlying mechanism for lipid elevations make a difference in the benefits that accrue.

Convincing evidence that modification of risk factors reduces disease incidence and mortality can be obtained only from experimental or clinical trials. Several primary prevention trials (selecting subjects free of disease at entry into the study) are under way to determine if altering diet, smoking, and controlling high blood pressure will lower the incidence of coronary heart disease. One such project, initiated in 1973, is called MRFIT, the Multiple Risk Factor Intervention Trial (Collaborating Investigators, 1976). It involves nearly 13,000 individuals at high risk of coronary heart disease, half of whom were randomly assigned to a special intervention program consisting of health education, behavior modification, group support approaches, and a maintenance program to prevent recidivism. The other half of the subjects received annual medical exams only. Another project, the Stanford Five Cities Program (Farquhar, 1978), is an extension of the first Stanford media-based intervention, with follow-ups being taken to determine heart disease morbidity and mortality.

Data regarding risk factor reduction in these two studies have not yet been published. It is not known how large a reduction in risk factors is necessary to observe a decrease in heart attacks in these populations. Results are expected to be available for the MRFIT in the next two years.

Clinical trials of life-style intervention involve the problems of behavioral measurement and of maintaining continued adherence to regimens (Kasl, 1980; Syme, 1978). Despite these disadvantages, such studies are major field trials of therapeutic and preventive measures that are relevant to the formation of public policy regarding behavior and health.

Concluding Comments

The biobehavioral approach to somatic health and illness is, by definition, an interdisciplinary venture. It requires the contributions of researchers representing a variety of skills and perspectives. Provision should be made for training investigators in the integrative skills necessary for continued progress in the scientific study of behavior and health and for providing them with appropriate research support.

The final section of this paper has highlighted the more promising research areas in behavior and health. Foremost among these are the study of psychosocial stress and the mechanisms linking stress and illness; psychoneuroimmunology; the challenge of maintaining abstinence from health-impairing behaviors; and techniques for enhancing medical compliance.

REFERENCES

Ader, R.
1976 "Psychosomatic research in animals." In C. Hill, ed., Modern Trends in Psychosomatic Medicine. London: Butterworth.

Ader, R., ed.
1981 Psychoneuroimmunology. New York: Academic Press.

Alexander, F.
1950 Psychosomatic Medicine. New York: Norton.

American Psychiatric Association
1968 Diagnostic and Statistical Manual of Mental Disorders (DSM-II). 2nd ed. Washington, D.C.: American Psychiatric Association.

Amkraut, A., and G. F. Solomon
1977 "From the symbolic stimulus to the pathophysiologic response: immune mechanisms." In S. J. Lipowski, D. R. Lipsitt, and P. C. Whybrow, eds., Psychosomatic Medicine: Current Trends and Clinical Applications. New York: Oxford University Press.

Baer, P. E., J. P. Vincent, B. J. Williams, G. G. Bourianuff, and P. C. Bartlett
1980 "Behavioral response to induced conflict in families with a hypertensive father." Hyptertension 2:170-177.

Bandura, A.
1977 Social Learning Theory. Englewood Cliffs, N.J.: Prentice-Hall.

Becker, M. H.
1979 "Understanding patient compliance: the contributions of attitudes and other psychological

factors." In S. J. Cohen, ed., New Directions in Patient Compliance. Lexington, Mass.: D. C. Heath.

Bernstein, D. A., and R. E. Glasgow
1979 "Smoking." In O. F. Pomerleau and J. P. Brady, eds., Behavioral Medicine: Theory and Practice. Baltimore: Williams and Wilkins.

Beecher, H. K.
1959 Measurement of Subjective Responses: Quantitative Effects of Drugs. New York: Oxford University Press.

Blumenthal, J. A., R. B. Williams, Y. Kong, S. M. Schanberg, and L. W. Thompson
1978 "Type A behavior and angiographically documented coronary disease." Circulation 58:634-639.

Breslow, L.
1978 "Risk factor intervention for health maintenance." Science 200:908-912.

Brooks, F. P.
1967 "Central neural control of acid secretion." In Handbook of Physiology, Section VI, Alimentary Canal. Baltimore: Williams and Wilkins.

Campbell, R. J., and J. P. Henry
1982 "Animal models of hyptertension." In D. S. Krantz, J. E. Singer, and A. Baum, eds., Handbook of Psychology and Health Volume 3: Cardiovascular Disorders and Behavior. Hillsdale, N.J.: Lawrence Erlbaum.

Cannon, W. B.
1942 "Voodoo death." American Anthropologist 44:169-181.

Cluff, L. E., A. Canter, and J. B. Inboden
1966 "Asian influenza: infection, disease, and psychological factors." Archives of Internal Medicine 117:159-163.

Cobb, S., and R. M. Rose
1973 "Hypertension, peptic ulcer, and diabetes in air traffic controllers." Journal of the American Medical Association 224:489-492.

Cohen, F., M. J. Horowitz, R. S. Lazarus, R. H. Moos, L. N. Robins, P. M. Rose, and M. Rutter
1980 Report of the Subpanel on Psychosocial Assets and Modifiers. Prepared for Committee to Study Research on Stress in Health and Disease, Institute of Medicine, National Academy of Sciences.

Cohen, S. J.
1979 New Directions in Patient Compliance. Lexington, Mass.: Lexington Books.

Cohen, S., G. W. Evans, D. S. Krantz, and D. Stokols
1980 "Physiological, motivational, and cognitive effects of aircraft noise on children." American Psychologist 35:231-243.

Collaborating Investigators
1976 "The multiple risk factor intervention trial (MRFIT)." Journal of the American Medical Association 234:825-828.

Cox, T.
1978 Stress. Baltimore: University Park Press.

Dahl, L. K., M. Heine, and L. Tassinari
1962 "Role of genetic factors in susceptibility to experimental hypertension due to chronic excess salt ingestion." Nature 194:480-482.

Dembroski, T. M., J. M. MacDougall, J. L. Shields, J. Petitto, and R. Lushene
1978 "Components of the Type A coronary-prone behavior pattern and cardiovascular responses to psychomotor performance challenge." Journal of Behavioral Medicine 1:159-176.

Dimsdale, J. E., T. P. Hackett, A. M. Hutter, and P. C. Block
1980 "The risk of Type A mediated coronary disease in different populations." Psychosomatic Medicine 42:55-62.

Dimsdale, J. E., and J. Moss
1980 "Plasma catecholamines in stress and exercise." Journal of the American Medical Association 243:340-342.

Dohrenwend, B. S., and B. P. Dohrenwend
1978 "Some issues in research on stressful life events." Journal of Nervous and Mental Disease 166:7-15.

Eisdorfer, C., and F. Wilkie
1977 "Stress, disease, aging, and behavior." In J. E. Birren and K. W. Schaie, eds., Handbook of the Psychology of Aging. New York: Van Nostrand Reinhold.

Engel, G. L.
1977 "The need for a new medical model: a challenge for biomedicine." Science 196:129-136.

Evans, R. I., R. M. Rozelle, S. E. Maxwell, B. E. Raines, C. A. Dill, T. J. Guthrie, A. H. Henderson, and P. C. Hill
1981 "Social modeling films to deter smoking in

adolescents: results of a three year field investigation." Journal of Applied Psychology 66:399-414.

Esler, M., S. Julius, A. Zweiflier, A. Randall, E. Harburg, H. Gardner, and V. DeQuattro
1977 "Mild high-renin essential hypertension: neurogenic human hypertension?" New England Journal of Medicine 296:405-411.

Falkner, B., G. Onesti, E. T. Angelakos, M. Fernandes, and C. Langman
1979 "Cardiovascular response to mental stress in normal adolescents with hypertensive parents. Hemodynamics and mental stress in adolescents." Hypertension 1:23-30.

Farquhar, J. W.
1978 "The community-based model of life style intervention trials." American Journal of Epidemiology 108:103-111.

Farquhar, J. W., N. Maccoby, P. D. Wood, et al.
1977 "Community education for cardiovascular health." Lancet 1:1192-1195.

Fordyce, W. E.
1976 Behavioral Methods for Chronic Pain and Illness. St. Louis: C. V. Mosby Co.

Fox, B. H.
1978 "Premorbid psychological factors as related to cancer incidence." Journal of Behavioral Medicine 1:45-133.

Frankenhaeuser, M.
1971 "Behavior and circulating catecholamines." Brain Research 31:241-262.

Fraumeni, J. F.
1975 Persons at High Risk of Cancer: An Approach to Cancer Etiology and Control. New York: Academic Press.

Friedman, M.
1979 "The modification of Type A behavior in post-infarction patients." American Heart Journal 97:551-560.

Friedman, M., R. H. Rosenman, and V. Carroll
1958 "Changes in the serum cholesterol and blood-clotting time in men subjected to cyclic variation of occupational stress." Circulation 17:852-861.

Friedman, M., J. H. Manwaring, R. H. Rosenman, G. Donlon, P. Ortega, and S. Grube
1973 "Instantaneous and sudden death: clinical and

pathological differentiation in coronary artery disease." Journal of the American Medical Association 225:1319-1328.

Friedman, R., and L. K. Dahl
1975 "The effects of chronic conflict on the blood pressure of rats with a genetic susceptibility to experimental hypertension." Psychosomatic Medicine 37:402-416.

Friedman, R., and J. Iwai
1976 "Genetic predisposition and stress-induced hypertension." Science 193:161-162.

Friedman, S. B., and L. A. Glasgow
1966 "Psychologic factors and resistance to infectious disease." Pediatric Clinics of North America 13:315-335.

Garfinkel, L.
1979 "Changes in the cigarette consumption of smokers in relation to changes in tar/nicotine content of cigarettes smoked." American Journal of Public Health 69:1274-1276.

Garrity, T. F., and M. B. Marx
1979 "Critical life events and coronary disease." In W. D. Gentry and R. B. Williams, Jr., eds., Psychological Aspects of Myocardial Infarction and Coronary Care. 2nd ed. St. Louis: C. V. Mosby.

Gatchel, R. J., and K. P. Price, eds.
1979 Clinical Applications of Biofeedback: Appraisal and Status. New York: Pergamon Press.

Glass, D. C.
1981 "Type A behavior: mechanisms linking behavioral and pathophysiologic processes." In J. Siegrist and M. J. Halhuber, eds., Myocardial Infarction and Psychosocial Risks. New York: Springer-Verlag.

Glass, D. C., and J. E. Singer
1972 Urban Stress: Experiments on Noise and Social Stressors. New York: Academic Press.

Glass, D. C., L. R. Krakoff, R. Contrada, W. C. Hilton, K. Kehoe, E. G. Mannucci, C. Collins, B. Snow, and E. Elting
1980 "Effect of harassment and competition upon cardiovascular and plasma catecholamine responses in Type A and Type B individuals." Psychophysiology 17:453-463.

Goldstein, J. L., and M. S. Brown
1974 "Binding and degradation of low density

lipoproteins by cultured human fibrobloasts." Journal of Biological Chemistry 249:5153-5162.

Graham, D. T.
1972 "Psychosomatic medicine." In N. S. Greenfield and R. A. Sternbach, eds., Handbook of Psychophysiology. New York: Holt, Rinehart and Winston.

Grunberg, N.E.
In Press The effects of nicotine on food consumption and taste preferences. Addictive Behaviors.

Hamburg, B. A., et al.
1980 "Executive summary." In B. A. Hamburg, L. F. Lipsett, G. E. Inoff, and A. L. Drash, eds., Behavioral and Psychosocial Issues in Diabetes: Proceedings of a National Conference. U.S. Department of Health and Human Services, Public Health Service Publication No. 80-1993. Washington, D.C.: U.S. Government Printing Office.

Harburg, E., J. C. Erfurt, L. S. Hauenstein, C. Chape, W. J. Schull, and M. A. Schork
1973 "Socio-ecological stress, suppressed hostility, skin color, and black-white male blood pressure: Detroit." Psychosomatic Medicine 35:276-296.

Harrell, J. P.
1980 "Psychological factors and hypertension: a status report." Psychological Bulletin 87: 482-501.

Haynes, S. G., M. Feinleib, and W. B. Kannel
1980 "The relationship of psychosocial factors to coronary heart disease in the Framingham Study. III. Eight-year incidence of coronary heart disease." American Journal of Epidemiology 3:37-58.

Henry, J. P., and J. C. Cassel
1969 "Psychosocial factors in essential hypertension: recent epidemiologic and animal experimental evidence." American Journal of Epidemiology 90:171.

Herd, A. J.
1978 "Physiological correlates of coronary-prone behavior." In T. M. Dembroski, S. M. Weiss, J. L. Shields, S. G. Haynes, and M. Feinleib, eds., Coronary-Prone Behavior. New York: Springer-Verlag.

Hinkle, L. E.

1974 "The effect of exposure to culture change, social change and changes in interpersonal relations on health." In B. S. Dohrenwend and B. P. Dohrenwend, eds., Stressful Life Events: Their Nature and Effects. New York: Wiley.

Holmes, T. H., and R. H. Rahe

1967 "The social readjustment rating scale." Journal of Psychosomatic Research 11:213-218.

House, J. S.

1975 "Occupational stress as a precursor to coronary disease." In W. D. Gentry and R. B. Williams, Jr., eds., Psychological Aspects of Myocardial Infarction and Coronary Care. St. Louis: C. V. Mosby.

Hunt, W. A., and J. D. Matarazzo

1970 "Habit mechanisms in smoking." In W. A. Hunt, ed., Learning Mechanisms on Smoking. Chicago: Aldine Publishing Co.

Hunt, W. A., J. D. Matarazzo, S. M. Weiss, and W. D. Gentry

1979 "Associative learning, habit and health behavior." Journal of Behavioral Medicine. 2:111-124.

Hurst, J. W., et al., eds.

1978 The Heart. New York: McGraw-Hill.

Institute of Medicine

1978 Perspectives on Health Formation and Disease Prevention in the United States. Report to National Academy of Sciences, Washington, D.C.

Jacobson, E.

1938 Progressive Relaxation. Chicago: University of Chicago Press.

Jarvik, M. E.

1979 "Biological influences on cigarette smoking." In Surgeon General's report, Smoking and Health. DHEW Publication No. 79-50066. Washington, D.C.: U.S. Government Printing Office.

Jenkins, C. D.

1971 "Psychologic and social precursors of coronary disease." New England Journal of Medicine 284:244-255, 307-317.

Julius, S., and M. Esler

1975 "Autonomic nervous cardiovascular regulation in borderline hypertension." American Journal of Cardiology 36:685-696.

Kannel, W. B., and T. R. Dawber
1971 "Hypertensive cardiovascular disease." In G. Onesti, K. E. Kim, and J. Hayer, eds., The Framingham Study, Hypertension: Mechanisms and Management. New York: Grune and Stratton.

Kannel, W. B., D. McGee, and T. Gordon
1976 "A general cardiovascular risk profile: the Framingham study." American Journal of Cardiology 38:46-51.

Kaplan, N. M.
1980 "The control of hypertension: a therapeutic breakthrough." American Scientist 68:537-545.

Kasl, S. V.
1980 "Cardiovascular risk reduction in a community setting: some comments." Journal of Consulting and Clinical Psychology 48:143-149.

Kasl, S. V., A. S. Evans, and J. C. Niederman
1979 "Psychosocial risk factors in the development of infectious mononucleosis." Psychosomatic Medicine 41:445-467.

Kawasaki, T., C. Delea, F. Bartter, and H. Smith
1978 "The effect of high-sodium and low-sodium intakes on blood pressure and other related variables in human subjects with idiopathic hypertension." American Journal of Medicine 64:193-198.

Korsch, B., and V. Negrete
1972 "Doctor-patient communication." Scientific American 227:66-78.

Krantz D. S.
1980 "Cognitive processes and recovery from heart attack: a review and theoretical analysis." Journal of Human Stress 6(3):27-38.

Krantz, D. S., M. E. Sanmarco, T. H. Selvester, and K. A. Mathews
1979 "Psychological correlates of progression of atherosclerosis in men." Psychosomatic Medicine 41:467-475.

Lazarus, A. S.
1966 Psychological Stress and the Coping Process. New York: McGraw-Hill.

Leventhal, H., and P. D. Cleary
1980 "The smoking problem: a review of the research and theory in behavioral risk modification." Psychological Bulletin 88:370-405.

Levi, L.
1979 "Psychosocial factors in preventive medicine." In Surgeon General's Background Papers for Healthy People Report. DHEW Publication #79-55011A. Washington, D.C.: U.S. Government Printing Office.

Levine, J. D., N. C. Gordon, and H. L. Fields
1978 "The mechanism of placebo analgesia." Lancet 2:654-657.

Ley, P., and M. S. Spelman
1967 Communicating with the Patient. London: Staples Press.

Liebeskind, J. C., and L. A. Paul
1977 "Psychological and physiological mechanisms of pain." Annual Review of Psychology 28:41-60.

Lown, B., R. Verrier, and R. Corbalan
1973 "Psychologic stress and threshold for repetitive ventricular response." Science 184:834-836.

Makidonan, T., and E. Yunis, eds.
1977 Immunology and Aging. New York: Plenum.

Mann, G. V.
1974 "The influence of obesity on health." New England Journal of Medicine 291:178-185, 226-232.
1977 "Diet-heart: end of an era." New England Journal of Medicine 297:644-650.

Manuck, S. B., and D. C. Schaefer
1978 "Stability of individual differences in cardiovascular reactivity." Physiology and Behavior 21:675-678.

Mason, J. W.
1971 "A re-evaluation of the concept of 'nonspecificity' in stress theory." Journal of Psychiatric Research 8:323-333.

Matarazzo, J. D.
1980 "Behavioral health and behavioral medicine: frontiers for a new health psychology." American Psychologist 35:807-817.

Mayer, D. J., D. D. Price, A. Rafii, and J. Barber
1976 "Acupuncture hypalgesia: evidence for activation of a central control system as a mechanism of action." In J. J. Bonica and D. Albe-Fesard, eds., Recent Advances in Pain Research and Therapy: Proceedings of the First World Congress on Pain. New York: Raven.

McAlister, A., P. Puska, K. Koskela, U. Pallonen, and N. Maccoby
1980 "Mass communication and community organization for public health education." American Psychologist 35:375-379.

McDonough, J. R., C. G. Hames, S. C. Stulb, et al.
1965 "Coronary heart disease among Negroes and whites in Evans County, Georgia." Journal of Chronic Disease 18:443-468.

Mechanic, D.
1968 Medical Sociology. New York: Free Press.

Medalie, J. H., B. Synder, J. J. Groen, H. N. Newfeld, U. Goldbourt, and E. Riss
1973 "Angina pectoris among 10,000 men: 5 year incidence and univariate analysis." American Journal of Medicine 55:583-594.

Melamed, B. G., and L. J. Siegal
1980 Behavioral Medicine: Practical Application in Health Care. New York: Springer Publishing Co.

Melzack, R., and P. D. Wall
1965 "Pain mechanisms: a new theory." Science 150: 971-979.

Meyer, A. J., J. D Nash, A. L. McAlister, N. Maccoby, and J. W. Farquhar
1980 "Skills training in a cardiovascular education campaign." Journal on Consulting and Clinical Psychology 48:129-142.

Meyer, R. J., and R. J. Haggerty
1962 "Streptococcal infections in families: factors altering individual susceptibility." Pediatrics 29:339-349.

Miller, N. E.
1969 "Learning of visceral and glandular responses." Science 163:434-445.
1976 "Behavioral medicine as a new frontier: opportunities and dangers." In S. M. Weiss, ed., Proceedings of the National Heart and Lung Institute Working Conference on Health Behavior. DHEW Publication No. (NIH) 76-868. Washington, D.C.: U.S. Government Printing Office.
1979 "General discussion and a review of recent results with paralyzed patients." In R. J. Gatchel and K. P. Price, eds., Clinical Applications of Biofeedback: Appraisal and Status. New York: Pergamon Press.

Miller, T., and J. S. Spratt
1979 "Critical review of reported psychological correlates of cancer prognosis and growth." In B. A. Stoll, ed., Mind and Cancer Prognosis. London: Wiley.

Nathan, P. W.
1976 "The gate-control theory of pain: a critical review." Brain 99:123-158.

National Science Foundation
1980 "Health of the American people." In Science and Technology: A Five-Year Outlook. Report from the National Academy of Sciences. Washington, D.C.: U.S. Government Printing Office.

Obrist, P. A.
1981 Cardiovascular Psychophysiology: A Perspective. New York: Plenum.

Osler, W.
1892 Lectures on Angina Pectoris and Allied States. New York: Appleton-Century-Crofts.

Page, I. H., and J. W. McCubbin
1966 "The physiology of arterial hypertension." In W. F Hamilton and P. Dow, eds., Handbook of Physiology: Circulation. Section 2, Volume 1. Washington, D.C.: American Physiological Society.

Pickering, G. W.
1967 "The inheritance of arterial pressure." In J. Stamler, R. Stamler, and T. N. Pullman, eds., The Epidemiology of Hypertension. New York: Grune and Stratton.

Pomerleau, O. F.
1979 "Why people smoke: current psychobiological models." In P. O. Davidson and S. M. Davidson, eds., Behavioral Medicine: Changing Health Life Styles. New York: Brunner-Mazel.

Pomerleau, O. F., and J. P. Brady
1979 Behavioral Medicine: Theory and Practice. Baltimore: Williams and Wilkins.

Puska, P., et al.
1978 Changing the Cardiovascular Risk in an Entire Community: The North Karelia Project. Paper presented at the International Symposium on Primary Prevention in Early Childhood or Atherosclerotic and Hypertensive Diseases, Chicago, October.

Ray, W. J., J. M. Racynski, T. Rogers, and W. H. Kimball, eds.
1979 Evaluation of Clinical Biofeedback. New York: Plenum.

Riley, V.
1975 Mouse mammary tumors: alteration of incidence as apparent functions of stress. Science 189: 465-467.

Riley, M. W., and A. Foner
1968 Aging and Society. New York: Russell Sage Foundation.

Riley, M. W., and B. A. Hamburg
1980 Report of the Subpanel on Stress, Health, and the Life Course. Prepared for the Committee to Study Research on Stress in Health and Disease, Institute of Medicine, National Academy of Sciences, Washington, D.C.

Rose, R. M., C. D. Jenkins, and M. W. Hurst
1978 Air traffic controllers health change study. FAA Contract No. DOT-FA73WA-3211, Boston University.

Rosenman, R. H., and M. Friedman
1974 "Neurogenic factors in pathogenesis of coronary heart disease." Medical Clinics of North America 58:269-279.

Rosenman, R. H., R. J. Brand, C. D. Jenkins, M. Friedman, R. Straus, and M. Wurm
1975 "Coronary heart disease in the Western Collaborative Group Study: final follow-up experience of 8 1/2 years." Journal of the American Medical Association 233:872-877.

Roskies, E.
1980 "Consideration in developing a treatment program for the coronary-prone (Type A) behavior pattern." In P. O. Davidson and S. M. Davidson, eds., Behavioral Medicine: Changing Health Lifestyles. New York: Brunner-Mazel.

Ross, R., and J. A. Glomset
1976 "The pathogenesis of atherosclerosis." New England Journal of Medicine 295:369-377, 420-425.

Russell, M. A. H.
1979 "Tobacco dependence: is nicotine rewarding or aversive?" Pp. 100-122 in N. A. Krasnegor, ed., Cigarette Smoking as a Dependence Process. NIDA Research Monograph 23. Alcohol, Drug Abuse and Mental Health Administration. DHEW Publication

No. (ADM) 79-800. Washington, D.C.: U.S. Department of Health, Education, and Welfare.

Sackett, D. L., and R. E. Haynes
1976 Compliance with Therapeutic Regimens. Baltimore: Johns Hopkins University Press.

Schachter, S.
1978 "Studies of the interaction of psychological and pharmacological determinants of smoking." Annals of Internal Medicine 88:104-114.

Schachter, S., B. Silverman, L. T. Kozlowski, D. Perlick, C. P. Herman, and B. Liebling
1977 "Studies of the interaction of psychological and pharmacological determinants of smoking." Journal of Experimental Psychology: General 106:3-40.

Schmale, A.
1981 "Stress and cancer." In Research on Stress in Health and Disease. National Academy of Sciences, Institute of Medicine, Washington, D.C.

Schneiderman, N.
1978 "Animal models relating behavioral stress and cardiovascular pathology." In T. M. Dembroski, S. M. Weiss, J. L. Shields, S. G. Haynes, and M. Feinleib, eds., Coronary-Prone Behavior. New York: Springer-Verlag.

Schwartz, G. E., A. P. Shapiro, D. P. Redmond, D. C. E. Ferguson, D. R. Ragland, and S. M. Weiss
1979 "Behavioral medicine approaches to hypertension: an integrative analysis of theory and research." Journal of Behavioral Medicine 2:311-364.

Selye, H.
1956 The Stress of Life. New York: McGraw-Hill.

Shapiro, A. K.
1971 "Placebo effects in medicine, psychotherapy, and psychoanalysis." In A. E. Bergin and S. L. Garfield, eds., Handbook of Psychotherapy and Behavioral Change. New York: Aldine.

Shapiro, A. P.
1982 "The non-pharmacologic treatment of hypertension." In D. S. Krantz, A. Baum, and J. E. Singer, eds., Handbook of Psychology and Health Volume 3: Cardiovascular Disorders and Behavior. Hillsdale, N.J.: Lawrence Erlbaum.

Siegel, S.
1979 "Pharmacological learning and drug dependence." In D. J. Oborne, M. M. Gruneberg, and J. R. Eiser, eds., Research in Psychology and Medicine. Volume II. London: Academic Press.

Stachnik, T.
1980 "Priorities for psychology in medical education and health care delivery." American Psychologist 35:8-15.

Svarstad, B. L.
1976 "Physician-patient communication and patient conformity with medical advice." In D. Mechanic, The Growth of Bureaucratic Medicine. New York: Wiley.

Syme, S. L.
1978 "Life style intervention in clinic-based trials." American Journal of Epidemiology 108:87-91.

Thomas, C. B., K. R. Duszynski, and J. W. Shaffer
1979 "Family attitudes reported in youth as potential predictors of cancer." Psychosomatic Medicine 41:287-302.

Turk, D. C., and M. Genest
1979 "Regulation of pain: the application of cognitive and behavioral techniques for prevention and remediation." In P. C. Kendall and S. D. Hollon, eds., Cognitive-Behavioral Interventions: Theory, Research and Procedures. New York: Academic Press.

U.S. Department of Health, Education, and Welfare
1964 Smoking and Health: A Report of the Surgeon General. U.S. Public Health Service Publication No. 1103. Washington, D.C.: U.S. Government Printing Office.
1978 Specialized Centers of Research in Arteriosclerosis: Cardiovascular Profile of 15,000 Children of School Age in Three Communities. U.S. Public Health Service Publication No. 78-1472. Washington, D.C.: U.S. Government Printing Office.
1979a Healthy People: A Report of the Surgeon General on Health Promotion and Disease Prevention. U.S. Public Health Service Publication No. 79-55071. Washington, D.C.: U.S. Government Printing Office.
1979b Smoking and Health: A Report of the Surgeon General. U.S. Public Health Service Publication No. 79-50066. Washington, D.C.: U.S. Government Printing Office.

Voors, A. W., T. A. Foster, R. R. Frerichs, L. S. Weber, and G. S. Berenson
1976 "Studies of blood pressure in children, ages

5-14 years, in a total biracial community." Circulation 54:319-327.

Weiner, H.
1977 Psychobiology and Human Disease. New York: Elsevier.

Weisenberg, M.
1977 "Pain and pain control." Psychological Bulletin 84:1008-1044.

Weiss, J. M.
1972 "Influence of psychological variables on stress-induced pathology." In R. Porter, ed., Physiology, Emotion, and Psychosomatic Illness: Ciba Foundation Symposium 8. Amsterdam: Associated Scientific Publishers.

Wolf, S., and H. G. Wolff
1947 Human Gastric Function. New York: Oxford.

Wolf, S., T. P. Almy, W. H. Backrach, H. M. Spiro, R. A. L. Sturdevant, and H. Weiner
1979 "The role of stress in peptic ulcer disease." Journal of Human Stress 5:27-37.

Earnings and the Distribution of Income: Insights from Economic Research

James J. Heckman and Robert T. Michael

INTRODUCTION

Few topics in the social sciences have received more study than the nature and causes of the distribution of income. In recent years, data, the statistical procedures for analyzing data, and the theoretical explanations for them have been the object of numerous studies.

This essay has two goals and those are reflected in the two parts of this paper. Part 1 presents salient characteristics of the distribution of income in the United States. Part 2 presents an outline of the main models that have been produced by economists to explain the distribution of wage earnings, the dominant component of income. It is hoped that the discussion in Part 2 will convey the essence of recent research on labor market behavior. The reader should be warned at the outset that there is less integration of the material in Parts 1 and 2 than one might wish. This reflects the fact that much more analysis remains to be done before the facts of the distribution of income are fully explained.

It is also important to stress at the outset that we say nothing about a topic of considerable interest to classical economists: the functional distribution of

---

This paper is a direct descendant of a literature review from a research proposal by the National Opinion Research Center. Due to publication deadlines, the revision of that material has been modest. We gratefully acknowledge the comments of many, especially John Abowd, Edward Lazear, and Robert Topel, in preparing the original manuscript, and we also thank Liz Peters and Marcia Weaver for research assistance.

income, i.e., its distribution among land, labor, and capital holders. In this paper we focus instead on the personal distribution of income, i.e., the distribution of labor earnings. We also ignore valuable work on aspects of the distribution of income considered in other social sciences, such as occupational mobility between generations and determinants of the size and type of governmental transfer payments.

1: THE NATURE OF THE DISTRIBUTION OF INCOME

Measures of Income Distribution

When one thinks of the distribution of income, there are a bewilderingly large number of individual characteristics that might be interesting or important to examine. What does the distribution look like by age? By education level? By ethnicity? By geographic region or state? Fortunately, the Current Population Reports (P-60 series) of the U.S. Bureau of the Census give detailed data for the United States based on current annual population surveys. One natural way to depict the distribution is illustrated in Table 1, which shows for 1978 the percentage of families in several specific income brackets. Each row in the table represents a distinct distribution. The distribution of money income before taxes and transfers is quite wide--some 8.2 percent of families have incomes under $5,000, while about 3.6 percent of families have incomes in excess of $50,000. Black families have considerably lower family incomes than white families, even when we look at the distributions for a specific education level of the household head.

Table 2 shows data in a different format. For selected years and types of households, each row in panel A shows what percentage of that row's income was received by each 20 percent of the population. For example, the top row tells us that in 1960 the bottom 20 percent of families received 4.8 percent of the income, while the middle 20 percent of families received 17.8 percent of the income and the top 20 percent received 41.3 percent of the income. Comparing across rows in Table 2 permits us to see how evenly or unevenly the income is distributed within each population group. Note how little change there seems to be between 1960 and 1978 in the distribution of income. We return to this observation below.

TABLE 1 The Distribution of Money Income in the United States, 1978 (Money Income Before Taxes and Transfers)

Family Type	No. of Families (millions)	Percentage Distribution of Families by Income Level ($ thousands)					
		Under $5.0	$5.0-6.9	...	$25.0-49.9	Over $50.0	Median Income
All	57.8	8.2%	6.0%	...	24.3%	3.6%	$17.6
White							
All	50.9	6.5	5.6	...	25.5	4.0	18.4
HH with 12 yr of school[a]	16.6	4.7	3.8	...	26.3	2.3	19.1
Black:							
All	5.9	22.4	9.8	...	12.8	0.6	10.9
HH with 12 yr of school	1.6	13.3	7.3	...	15.3	0.3	13.9
Spanish origin:[b]							
All	2.7	14.6	9.2	...	12.5	1.0	12.6
HH with 12 yr of school	0.6	8.0	7.8	...	17.1	0.9	15.7

[a] HH is householder.
[b] Persons of Spanish origin may be of any race.

SOURCE: U.S. Statistical Abstract (1980: Table 748, p. 452).

The Lorenz curve, one of the standard tools for looking at the distribution of income, uses information in a form quite similar to Table 2, which shows the percentage of households (from poorest to wealthiest) and the percentage of total income received by those households; the Lorenz curve was devised around 1910. Figure 1 is a typical Lorenz curve for the United States in a recent year. It shows, as does Table 2, that the lower half of households gets about one-quarter of the total income and the lower three-quarters of households gets about half the total income. Notice that if the Lorenz curve were less bowed, the income would be more evenly distributed. The straight line in the figure in fact would characterize a completely equal distribution of income. The lowest 10 percent of households would get 10 percent of the income and so would the upper end--indeed, there would not be a lower or upper end in that case. Likewise, if the curve were more bowed, it would reflect even greater unevenness in the distribution. This property of the Lorenz curve can be used to produce a convenient summary statistic that reflects how

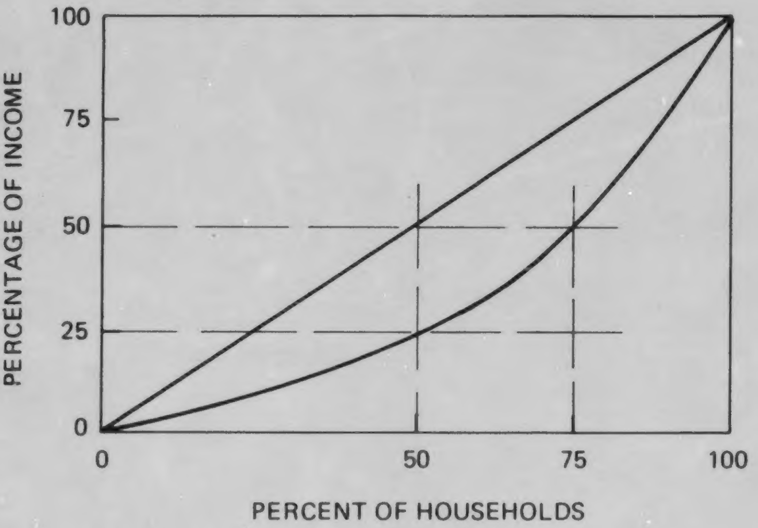

FIGURE 1 Lorenz curve.

TABLE 2 Characteristics of the Distribution of Income in the United States, for Selected Years and Groups

	Lowest Fifth	Second Fifth	Third Fifth	Fourth Fifth	Highest Fifth	(Highest 5%)
A: Percentage of Aggregate Money Income of Families by Quintile						
All families						
1960	4.8	12.2	17.8	24.0	41.3	(15.9)
1970	5.4	12.2	17.6	23.8	40.9	(15.6)
1978	5.2	11.6	17.5	24.1	41.5	(15.6)
White families						
1960	5.2	12.7	17.8	23.7	40.7	(15.7)
1978	5.6	12.0	17.6	23.9	41.0	(15.5)

Black and other families						
1960	3.7	9.7	16.5	25.2	44.9	(16.2)
1978	4.2	9.6	16.3	25.1	44.7	(15.9)
Unrelated individuals						
1960	1.7	7.3	13.7	26.0	51.4	(20.2)
1978	4.1	9.0	14.9	23.9	48.2	(19.5)

B: Income Level (in 1978 dollars, $ thousands) at Upper End of Each Quintile

All families						
1960	6.1	10.6	14.0	19.4	--	--
1978	8.7	14.7	20.6	28.6		
Unrelated Individuals						
1960	1.4	2.6	5.3	9.2	--	--
1978	3.0	5.1	8.3	13.3		

SOURCE: U.S. Statistical Abstract (1980: Table 752, p. 454).

evenly income is distributed. If we form the ratio of the area inside the bowed-out curve to the area in the whole lower triangle in the figure, that ratio would be close to zero when incomes are almost evenly distributed. The ratio would be close to 1.0 if income were almost completely unevenly distributed (e.g., if one household got almost all the income). This ratio, called the concentration ratio or the Gini coefficient can range from 0 to 1.0; the bigger it is, the more unevenness or inequality there is in the income distribution. The Gini coefficient for income among families and unrelated individuals in the United States is about 0.40. We will discuss that number shortly.

Before turning to more detail about the distribution, it is important to describe another, rather different measure of the dispersion in income. Generally speaking, the distribution of income when drawn as a frequency distribution as in Figure 2 has a particular shape with a noticeably longer and thicker upper than lower tail. The distribution is not symmetric on either side of its mean value; it has a "positive skew." It is often asserted that the amount of asymmetry in the income distribution is such that a distribution of the log of income is nearly symmetrical and shaped very nearly like a normal distribution. A normal distribution can be characterized completely by its mean and its variance, so the variance of the log of income is often used as a measure of the dispersion or inequality in income (for other measures of income inequality and their attributes, see Kakwani, 1980).

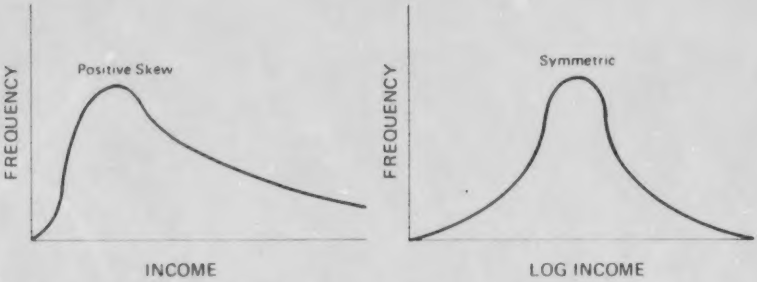

FIGURE 2 Frequency distribution of income.

Apparent Income Stability Amid Social Change

The Evident Stability

As suggested by the first three rows of Table 2 and confirmed time and again in the literature, the overall distribution of money income in the United States shows remarkable stability over the past few decades. A Gini coefficient for income among families and unrelated individuals by year is measured as .41 in 1951 and as .41 in 1974. Not only was there no apparent secular trend in the Gini coefficient over this quarter century, but also the year-to-year fluctuations were trivial (the annual Gini coefficient ranged from .40 to .42; Treas and Walther, 1978). A similar conclusion is reached using another measure of the distribution of income, the income received by the lowest 20 percent of the income distribution among families. That percentage was 5.1 percent of the before-tax aggregate income in the late 1940s and remained at about 5 percent throughout the subsequent period, as shown in Table 2. The stability in the income received by the lowest quintile is mirrored in the income received by the other quintiles throughout the distribution.

The perception of no change in the inequality of income in the United States over the past 25 years or so is somewhat modified when the distribution is characterized by the variance of the log of earnings. The variance has declined somewhat since 1950. For example, the variance of the log of income among men ages 25-64 in the United States fell from 0.65 in 1949 to 0.58 in 1969 (Chiswick and Mincer, 1972:560). The discrepancies in trends between the Gini coefficient and the variance of the log of earnings reflect the fact that the Gini tends to emphasize changes around the center of the distribution, whereas the variance measure is relatively more sensitive to changes in the lower tail of the distribution (see Atkinson, 1970).

The general impression from these measures of economic inequality is one of relative stability. This conclusion holds if one looks at the distribution of money income adjusted for taxes and transfers. Reynolds and Smolensky (1977) employ a concept of net national product as a measure of income and attempt to identify the impact of government transfers and taxes on the distribution of income from 1950 through 1970. Their assessment suggests that net income (after taxes and transfers) is substantially more evenly distributed than gross income, but that the distributions of gross and net income do not exhibit any appreciable trend over the 20 years of their study. For example, on the basis of one of their assumptions Reynolds and Smolensky report that the Gini coefficient for net income was .315 in 1950 and .290 by 1970 (Reynolds and Smolensky, 1977:71).

The assessment of the impact of government tax and transfer programs has been undertaken in several studies, and the numbers obtained vary depending on the set of assumptions used to produce the estimates (for example, the assumptions about the distribution of real benefits from the general expenditures of the federal government). Much of this literature considers the net effect of government taxes and transfers on the shape of the distribution (e.g., Golladay and Haveman, 1977; Pechman and Okner, 1974; Reynolds and Smolensky, 1977; for a review, see Danziger et al., 1981), but in terms of trends over time, the Reynolds and Smolensky conclusion of no significant trend is the current consensus in this literature. Thus the evidence suggests remarkable stability in the distribution of income over the past 25 years, as gauged by any of the conventional measures and defined as narrowly as earnings or as broadly as income after taxes and transfers.

This stability is quite surprising. Fundamental changes in the social, economic, and political structure of the United States have occurred during the past 30 years. The social changes include substantial fluctuations in fertility rates (in particular the baby boom of the 1950s and the subsequent decline in fertility through the 1960s and most of the 1970s), the recent increase in divorce rates and rising age at first marriage, and declining mortality rates. As a result, the nature of the household structure in the United States, the age structure of the population (including the "dependency ratio"), and the age distribution of the work force are profoundly different in 1980 from what they were in 1960.

The important economic changes over the past two or three decades include substantial growth in real income, until the recent slump in productivity growth; changes in the composition of the work force, reflecting the increased participation of women in the labor market, the trend toward earlier retirement, and the maturing of the baby boom cohort; the continued shift toward the service sector of the economy; and, within the past decade, changes in the relative price of sources of energy.

Finally, evidence of changing political climate within the United States is the growth of government and an increasing reliance on government to influence the nature, quality, and distribution of products in the American economy, as well as to shape the social environment in which we live. Domestic policy has gone from what might be characterized as limited involvement in the 1950s through a period of active intervention in civil rights and the war on poverty in the 1960s to the recent conservatism of the Reagan era. This political cycle has been accompanied by a shift in attitude within the body politic from apathy in the 1950s to optimism about the efficacy of social policy instruments in the 1960s to general skepticism, if not pessimism, about the role of government in recent years.

In the face of these structural changes, the stability in the inequality of the overall distribution of money income in the United States is remarkable. As the next section suggests, the apparent stability masks much change.

Offsetting Factors: Age and Household Structure

Several studies have attempted to probe this global stability in the income distribution and, in doing so, have identified offsetting forces. Two factors that have been shown to play a fundamental role in the apparent stability of the income distribution in recent decades are the distributions of the population by age and by household structure.

Paglin (1975) examined the effects of age on the distribution of income and distinguished two effects. The first is that income tends to increase with age, and it so happens that the rate of growth of income with age has increased over the postwar period. (The human capital model discussed below provides a reason for the growth in income with age.) For a particular population age distribution, Paglin's first effect increases income inequality over time solely because income-age profiles are becoming steeper. Paglin's second effect is that the population distribution by age has changed over the postwar period in such a way that for a particular age-income profile, the distribution of income has become more unequal. So both of these effects increase the inequality in income over time. Yet, as we stressed above, the aggregate income distribution shows a virtual constancy over time. So there must be an offsetting effect that has decreased inequality. The apparent stability masks a trend toward equality offset by the changing age distribution of the population. (While Paglin's general point is now accepted, reservations about his methodology have been voiced; see Pyatt, 1976, and, for related discussions, Blinder, 1980; Danziger et al., 1977. Also see Lillard, 1977, for an alternative method of age adjustment.)

Regarding adjustments for family structure, in an important study Treas and Walther (1978) show dramatically yet simply the relevance of family structure for the distribution of income. With data from the Current Population Survey, Treas and Walther show that within practically every category of household structure--female-headed families, husband-wife families with wife employed, husband-wife families with wife not employed, unrelated women, and unrelated men--there was a decline in inequality of income over the period from 1951 through 1974. For most of the categories the decline in inequality was a statistically significant one and substantial in magnitude. The measured inequality over all categories combined, however, shows no trend over the period. The reconciliation of the paradoxical finding of no overall trend yet a substantial trend toward equality within practically every component of the total involves shifts in the weights of the various categories through time. Household categories characterized by especially low incomes rose as a proportion of the total population of households. These categories include unrelated women, unrelated men, and households headed by women. Households with the highest average household income--those in which both husband and wife were employed--also became proportionately more numerous. On the other hand, the number of households with incomes near the overall mean, that is, husband-wife households in which only the husband was employed--declined as a proportion of the total. Thus the overall stability of the aggregate distribution of household income masks a shift toward the tails of the distribution in the types of household structures in the population, offset by substantially less inequality within each household category.

The stability in inequality over time in earnings among men of prime working age is not found in the distribution of earnings among all workers, which has shown an increase in inequality over time, or in the distribution of income among families and unrelated individuals, which has shown a decrease in inequality over time. These differences are explained by and stress the importance of the unit of analysis. Among all workers the influx of teenagers and women has resulted in an increase in the dispersion of earnings. When viewed from the perspective of the family unit, the increased labor force participation of married women has tended to reduce the dispersion of income among families.

Trends for Specific Groups

Turning to measures of inequality over time for specific groups, Table 3 shows the variance of the log of income among persons ages 14 and over with income, by sex and age, for selected years from 1950 to 1970, expressed relative to the variance for men of prime working age in 1950. For men (the top panel of the table) inequality of income is relatively great at young and old ages compared with middle ages; for women that tendency is less pronounced. Over time, inequality increases for all groups except elderly men. Over all groups combined and for each sex separately there has been substantial increase in inequality, reflecting a steepening age earnings profile. Schultz (1975) among others stresses that the increased inequality in income or earnings among earners is in some degree a reflection of the growing tendency toward part-time employment. The growth in part-time employment among the young, the elderly, and married women tends to increase inequality measured among persons with income.

Regarding black-white wage differentials, the postwar era has witnessed a trend toward equality. As reported in Smith and Welch (1978:3), the ratio of wage and salary earnings of black men to those of white men rose substantially from .54 in 1947 to .58 in 1959 to .67 in 1969 to .73 in 1975. Among full-time workers a similar trend is observed: The ratio was at .64 in 1947, suffered a cyclic decline to .61 in 1959, then rose to .69 in 1969 and to .77 in 1975. The relative gains for black women were more substantial than those for black men. Based on data from the _Current Population Survey_, Freeman (1978) reports the ratio of median wage and salary of black to white _men_ rising from .50 to .73 from 1950 to 1975 and for black to white _women_ from .40 to .97 over that period. So black women had achieved near parity with white women in earnings by 1979. A major contributing factor to the relative improvements of blacks has been their relative gains in education. The black:white ratio of mean schooling level of new labor market entrants rose from .72 in 1940 to .82

TABLE 3 Relative Inequality in Personal Income (Among Persons 14+ with Income) by Year, Age, and Sex: Estimates of Variance of Logs, Expressed Relative to the Variance for Men Ages 35-44 in 1950

	1950	1955	1960	1965	1970
Men					
Total	165	187	217	232	252
14-19	144	208	224	201	220
20-24	109	112	146	166	175
25-34	80	94	91	90	97
35-44	100	99	118	115	102
45-54	136	147	153	136	124
55-64	169	170	183	175	163
65+	211	179	151	134	136
Women					
Total	165	192	213	232	248
14-19	136	138	260	273	220
20-24	118	148	182	204	198
25-34	146	174	213	224	252
35-44	157	188	206	214	229
45-54	91	181	204	205	221
55-64	171	198	204	222	233
65+	118	117	111	110	125
Men and women	199	230	266	277	299

Note: Index: 100 = men ages 35-44 in 1950 whose variance was 0.4709.

SOURCE: Derived from Table 6-2, p. 154 (Schultz, 1975) from P-60 Series CPR data.

in 1950 to .88 in 1960 to .99 in 1970 (Smith and Welch, 1978).

Regarding the wage differential by sex, neither the trend toward wage equality in the postwar era nor the interpretation of the generally lower wage of women is as evident as for the black-white wage differential. While the unadjusted annual earnings differential for women relative to men actually declined from 1949 to 1959 to 1969 from .56 to .50 to .47, that decline is explained by the fact that the average hours and weeks worked by women also fell over that period as more and more women entered the labor force as part-time workers. Adjusted for hours of work, the corresponding figures were .67 to .66 to .63, and adjustment for the difference in education levels further alters the picture to a trendless differential of .63 to .65 to .63 over that same 1949-1969 period (see

Chiswick and O'Neill, 1977:28). When Fuchs (1974) performed comparable adjustments for schooling level for whites only, he found hourly earnings had actually risen for women relative to men from 1959 to 1969 by as much as 11 percent for those with 12 or more years of schooling and had risen as well when age-specific groups were compared. These gains in wage differentials for white women compared with white men were striking in light of the rapidly growing female employment that introduced relatively inexperienced women into the labor market.

Income Mobility

The Extent of Mobility

Perhaps as important as the dispersion of income at any single time is the degree of mobility within that distribution over time. Data show that the distribution of income is characterized by considerable period-to-period mobility. Likewise, studies of mobility from generation to generation demonstrate considerable mobility in income and occupational status. Studies in the economic literature find substantial mobility in the United States income distribution, especially for the young and for those near the bottom of the distribution (see Mirer, 1975; Benus, 1974; Kohen, Parnes, and Shea, 1975; Blinder, 1980).

Mobility from Poverty

Whereas research on occupational and income mobility across generations examines the entire income distribution, research on income mobility from period to period has frequently focused on income mobility from poverty to nonpoverty. This research has served to emphasize a distinction between those for whom poverty is a relatively transitory phenomenon and those for whom poverty is a more permanent circumstance. Whether poverty is seen as a transitory or a permanent phenomenon depends in part on the time interval over which a study focuses. For example, two studies with different data sets and definitions of poverty show a similar incidence of poverty in one year conditional on being in poverty the previous year. The 1976 report, The Measure of Poverty, using data from the Denver Income Maintenance Experiment, found that in 1970 50 percent of those in poverty remained in poverty 12 months later (U.S. Department of Health, Education, and Welfare, 1976). Lillard and Willis (1978), using data from the University of Michigan Panel Study of Income Dynamics, reported that the probability of a man in poverty in 1967 being in poverty the following year was 34 percent for whites and 61 percent for blacks (p. 1004). The former study, however, looks at intrayear patterns and finds little mobility (i.e., at most, 9 percent changed status from poverty to nonpoverty during any one month in 1970), while the latter study looks at interyear patterns and finds substantial mobility (i.e., the probability that a man in poverty in 1967 would be in poverty in 1973 was 28 percent for whites and 46 percent for blacks). Of course, the extent to which measurement error in the income affects this evidence of mobility over time deserves careful scrutiny.

Given the distinction between those for whom poverty is a permanent status and those for whom it is transitory, the determinants of mobility become important. Lillard and Willis conclude that variables such as race, schooling, and job experience are important determinants of permanent poverty status. Other variables affecting the mobility of families have also been examined. The recent increase in the number of earners in multiperson families has reduced the number of family units in poverty. Levy (1976) studied changes in family structure to examine the effects of increases in female-headed families over the last few years.

The consensus of research on mobility can be expressed by quoting from Aaron's thoughtful synthesis of recent policy-related research (1978:49):

> The number of people who are sometimes poor is far larger and the number who are always poor is at least somewhat smaller than official statistics suggest: many of the poor differ from the rest of us only in their lack of money, and many of them one day will leave poverty. But many of the rest of us one day will be poor too.

Poverty

Underlying much of the discussion of income mobility between poverty and nonpoverty are questions about the definition and measurement of poverty. The history of the official poverty definition, the alternative definitions, and the limitations in the measures used are quite well explored in the excellent summary provided in the congressionally mandated Report on Poverty (Mahoney, 1976) and its technical supplement (especially Vol. I: Urban Systems Research and Engineering, 1976). The official measure of poverty combines a political decision about the appropriate level of income at which to dichotomize households as either in poverty or not in poverty and an administrative ruling about the interpretation of that decision for various family sizes and structures. The advantages and disadvantages of measures of poverty based on absolute or relative standards have been well reviewed (for example, see Blinder, 1980; Urban Systems Research and Engineering, 1976).

Regarding trends in the incidence of poverty, conclusions depend on the definitions of poverty and of income. Reflecting the overall stability in the distribution of income over time discussed above, for poverty measured in relative terms there has been no discernible trend in the incidence of poverty. Using a definition of relative poverty Plotnick and Skidmore (1975, ch. 4) show no movement between 1965 and 1972; Mahoney (1976, ch. 5) shows no substantial trend between 1967 and 1974 when poverty is measured as 50 percent of U.S. median income. Measurements using a definition of absolute poverty (that make adjustment for inflation), however, show a decline over the past 20 years in the proportion of the population in poverty and in the number of persons in poverty. Based on an income measure adjusted for cash transfers, Plotnick and Skidmore (1975) report a decline in the number of poor persons from 38 million in 1962 to an estimated 27 million by 1975 (p. 82) and, as a percentage of the total population, a decline from about 21 percent in 1962 to about 11 percent by 1973 (p. 84). The 11 percent appears to have remained steady through 1978 (Statistical Abstract of the United States 1980:464).

When in-kind transfers in the form of medical care, food, and housing are included, the incidence of poverty is considerably less and the decline considerably more dramatic. From figures on income excluding in-kind transfers roughly comparable to those cited from the Plotnick and Skidmore study, Paglin (1977) estimates that in-kind transfers have reduced the incidence of poverty over the same time period (1962-1975) from about 30 million persons to a little over 6 million persons, or from about 16 percent of the population in 1962 to only about 4 percent in 1973 (1977, Table 8). It should be noted, however, that a large portion of the in-kind assistance is in the form of health care through the Medicaid and Medicare programs, and there is substantial controversy about whether that transfer should be counted as income. For example, a Congressional Budget Office (CBO) study estimated that in 1976, 9.3 percent of all families were below the poverty level without counting Medicaid and Medicare, whereas only 6.7 percent were below the poverty level when these benefits were counted (Congressional Budget Office, 1977:7-9).

Interpretation of the effects of in-kind transfers on poverty incidence is controversial in general, because of problems that arise in evaluating the usefulness of the in-kind transfer to the recipient and determining how much of the aggregate transfer is received by persons in poverty. There is at least consensus that including in-kind transfers appreciably reduces the incidence of poverty and that the downward trend in the poverty population is considerably stronger when in-kind as well as cash transfers are taken into account.

The incidence of poverty measured by pretransfer income also varies over time because of the differential effects on different groups of macroeconomic conditions such as unemployment. For example, Plotnick and Skidmore (1975) report that during the 1965-1972 period of overall full employment and economic growth, poverty incidence fell for most demographic subgroups but rose for families headed by women. In contrast, during the 1972-1975 period of higher unemployment, poverty incidence shifted toward the working poor as relatively more families headed by men of prime working age fell below the poverty line.

We turn briefly to the incidence of poverty among specific groups. In the mid-1970s, the incidence of poverty was particularly great for blacks, the elderly, and persons living in female-headed families: In 1974, blacks composed about 12 percent of the U.S. population but 31 percent of the poverty population; the elderly constituted 10 percent of the U.S. population but 14 percent of the poverty population; and persons living in female-headed families accounted for 11 percent of the U.S. population but 35 percent of the poverty population (Mahoney, 1976:112).

CBO figures emphasize in another way both the impact of transfers on the incidence of poverty and the differential incidence by age. For 1976 CBO estimates that, considering pretax, pretransfer income, 18.6 percent of families headed by someone under age 65 and 59.9 percent of families headed by someone over age 65 were in poverty.

When taxes, social insurance income, cash, and in-kind transfers are taken into account, those percentages fall, respectively, to 8.9 percent and 6.1 percent (Congressional Budget Office, 1977:12). It is only in the context of these latter, low figures that one might argue, as has Martin Anderson (1978:37), that "the 'war on poverty' has been won . . . except for perhaps a few mopping-up operations."

One final point deserves emphasis. A conclusion of perhaps greater relevance than either the precise incidence of poverty today or the trend in it over the past decade is the finding that there has been little change in the aggregate poverty rate in the United States based on measures of income exclusive of transfers (Danzinger et al., 1979:2, 30):

> If only earned income is considered, the aggregate incidence of poverty has remained unchanged at about 21 percent. . . . The level of absolute pretransfer poverty has been stagnant since 1965. In both 1976 and 1965, 21 percent of all persons were pretransfer poor. Although the incidence . . . was slightly lower in the intervening years, there is no downward trend.

The evidence clearly indicates that the measured reduction in poverty has resulted from tranfers of income rather than employment. The interpretation of this evidence is not without ambiguity. Income-conditioned transfers create a disincentive to work for those in the lower tail of the pretransfer earnings distribution and hence would tend to increase inequality in earnings. The fact that there is no increase in earnings inequality is consistent with two interpretations: (1) that income transfers did not reduce work effort or (2) that income transfers did reduce work effort but were somehow offset. The social, economic, and political consequences of using transfer income rather than jobs as the main mechanism for the reduction of poverty in the United States surely warrants much further exploration.

Throughout this section we have discussed the distribution of money income, sometimes wage earnings and sometimes total money income. There may be, of course, important distinctions between the distribution of money income and the distribution of welfare for at least three reasons.

First, the income recipient unit, e.g., the family, is not homogeneous across all units and so income per unit is not an unambiguous measure of welfare. Family size differences make income per family not comparable across families. Dividing by the number of family members creates a per capita income measure, but studies of scale economies with family size suggest that per capita measures are an overadjustment (see Lazear-Michael, 1980; Mahoney, 1971; Muellbauer, 1977). The equivalence of income across families of different sizes is an important issue of current research concern. Likewise, family structure--single parent versus two-parent families and one-earner versus two-earner families--raises similar issues of income equivalence.

Second, money income does not include all aspects of real income. Differences in environment from the cleanliness of the air to political stability and to the geological and social climates in which people live are in a broader definition a part of income. Amenities of job, neighborhood, and city might be included as well in measures of real income, as reflected in recent efforts to quantify "social indicators." Differences in hours of work create observed differences in money income that overstate inequality by disregarding the offsetting differences in leisure (or nonwork) time.

Third, and most imponderably, the translation of any given income level into welfare may differ across individuals because of differences in the utility functions of individuals.

2: THE DETERMINANTS OF EARNINGS

Despite the reduced incidence of absolute poverty in the United States, there has been almost no reduction in the poverty rate based on income exclusive of transfers. That is, the considerable success of the war on poverty has been the result of income transfers, not the result of growth in earnings among those previously in poverty. There have been great strides forward, however, in the reduction of measured wage differentials between minority and majority groups and in providing equality of educational opportunity and, thus, educational achievement. These and other gains have apparently not had an impact on labor market earnings sufficient to allow them to replace transfer income as a major factor in reducing absolute poverty.

More than two-thirds of income comes in the form of earnings. For this reason, and given the considerable research progress in this area in recent years, we focus in this section on the determinants of earnings. We summarize the current understanding of the workings of the labor market and of the supply and demand for labor services. As noted in the introduction, this discussion is not specifically directed at the features of the distribution of income, as in the preceding section.

Conventional Neoclassical Models and Their Extensions

The Wage Rate

Labor market income is the product of two distinct factors: wage rates, which measure earnings per unit of labor, and labor supply. It is impossible to disentangle the two concepts completely, since labor supply decisions made today can affect the wage in the future and expected future wage rates affect current and future labor supply. Nevertheless, it is analytically convenient to distinguish the topics of wage rate determination and labor supply determination.

The most important current theory of the generation of earning power is the human capital model, pioneered some 20 years ago by Becker (1960, 1975), Mincer (1958, 1962a), and Schultz (1961), with antecedents as far back as Adam Smith, the father of modern economics. The best review of the work on this topic through the mid-1970s is Rosen (1977). Other systematic reviews of this literature include Kiker (1971), Blaug (1976), and, with special reference to income distribution, Mincer (1970).

The theory of human capital is better characterized as a broad idea than a precisely delineated model: An individual spends current resources (both direct expenditures of dollars and foregone earnings) in order to raise his or her productivity or wages in the future. That expenditure constitutes the production of a capital asset embedded in the individual--that is, his or her human capital. It may be an investment in formal schooling, in job training, in one's health, in one's location (by migration), in information about available jobs (job search), or in providing information about one's skills (advertising or signaling). The payoff may come in areas outside the labor market as well as within it. (The vast proportion of the literature considers the labor market returns to these investments, but see Michael, 1982, for a review of the benefits of schooling outside the labor market.)

The theory of human capital investment focused the attention of labor earnings research on the estimation of the relationship between wages and a wide variety of productivity-related characteristics, such as formal schooling, labor force experience, age, ability measures, family background measures, and many others. Becker (1975) provided quantitative evidence on the effect of schooling on earnings in the form of a rate of return on the investment, and now empirical estimates of the rate of return to schooling abound.

Griliches (1975) summarizes the empirical evidence on the relationship between earnings and schooling. He also reviews the many correlations that have been made to avoid "ability bias"--the bias that arises in estimating the effect of schooling on earnings when the unobservable components of earnings potential are correlated with measured education components. (Thus the question arises whether education is merely a proxy for these unobservables.) He concludes that a positive association remains between the two when ability bias is eliminated by econometric procedures.

Ability bias is a special case of a general problem. When one estimates the effects of schooling or job experience on the wage rate, there are typically many additional unobserved factors that may affect the estimated relationship. These unobserved factors may include, in addition to ability, aspects of family background, early childhood development, genetic endowments, school performance, school quality, educational financial opportunities, and many details of both the educational and labor force histories of the worker, not to mention health, motivation, energy level, psychological characteristics, and so on. Studies have used a remarkably wide range of variables from large cross-sectional data sets to attempt to measure and adjust for effects of ability, family background, school characteristics, health, and even genetic endowments on labor earnings. The empirical importance of background variables is difficult to summarize and remains a matter of dispute.

In virtually all studies, schooling has a positive measured effect on earnings. This effect does not appear to be greatly changed when a variety of methods are used to control for unobservable background variables. Taubman (1976) presents a minority view arguing that there is substantial bias in the conventional estimates of the effect of schooling on earnings. Taubman uses many variables as controls in his earnings function. As noted by Welch (1975), however, the proliferation of control variables is itself likely to cause a severe downward bias in the estimated effect of schooling on earnings. (This follows because Welch assumes that schooling is only a proxy for true human capital.)

Goldberger (1975), in his review of heritability, demonstrates that there is considerable controversy in this literature about the relative strength of genetic and environmental effects in determining the productive capacities of individuals. Goldberger shows that attempts to disentangle heredity from environment have typically been based on arbitrary assumptions.

Independent evidence from time-budget studies documents the importance of parental time in child care on the subsequent schooling performance of children (Hill and Stafford, 1974; Leibowitz, 1974). This finding is an important link in understanding the mechanism through which family background variables operate on earnings potential.

New econometric methods, notably techniques developed by Chamberlain and Griliches, address the important issue that individuals are not assigned into schooling categories at random, so one cannot treat the schooling-earnings relationship as equivalent to a controlled experiment. The schooling decision itself is important; it sorts workers into groups that are, obviously, not random samples of the entire population. One should not look at the increased earnings of those with more schooling and draw some implication for social policy about giving people an extra dose of schooling. Such implications will be misguided unless the nonrandomness of the survey data--the endogenous nature of the schooling decision--is taken into account. This line of work is one of the most important currently under way in this area. (See also work by Willis and Rosen, 1979, who address this problem in a "sample selection bias" setting.)

Several other measurement problems have been addressed in the human capital literature. Since a variety of human capital investments may be pursued by an individual at any time, it is often not possible to identify empirically the costs and returns of specific types of human capital investments without making specific assumptions about the important sources of investment and the form of the investment functions. The required extra information is frequently provided by ignoring all other components of human capital and concentrating on the effects of schooling or postschool training on the growth in wages over an individual's career. Ben-Porath (1967) presented the first rigorous model of human capital accumulation. His model has since been generalized by others and estimated in papers that explore the dynamics of the evolution of earnings and provide a more firm statistical foundation for the human capital earnings model (see Haley, 1976; Heckman, 1976; Lillard, 1977; Rosen, 1976).

This work is still at a relatively primitive stage and has not received much attention since the mid-1970s (but see Killingsworth, 1982). These investment models are potentially important guides to general policy analysis, since they indicate how individuals choose remuneration packages with low wages and large training components early in their careers in anticipation of greater wage growth over the life cycle. In addition, since the wages actually received during the investment period are lower than the wages workers could receive at that point in their life cycle, the human capital model provides an explanation for the divergence of observed wages. A key implication of this divergence is that labor income is an incomplete measure of total employment compensation (since workers may also be acquiring new skills that will pay off later) and may therefore be an inappropriate policy target by itself. Empirical results (e.g., reported by Heckman, 1976; Johnson, 1975; Lazear, 1976; and Mincer, 1974) indicate that a substantial proportion of a typical white male worker's early career is spent investing in human capital. Lazear (1979a) uses this type of argument in the study of wage differentials by sex. He suggests that if young women have recently been receiving substantially more on-the-job training than previously, adding the value of this training to the pecuniary component of the wage earned would further raise the rate of growth of the wages of young women relative to young men.

The advantage of more explicit and focused human capital models lies in the identification of parameters that can be used to evaluate the effects of alternative policies. The computational complexity of the structural models has slowed their adoption by applied economists. A majority of the empirical studies of wage determination employ only a descriptive model instead of a structurally identified model.

A recurring problem of measurement is the bias introduced by using a cross-sectional survey in place of longitudinal data on cohorts to estimate age or experience effects on wage changes over a lifetime. Theoretical studies of earnings equations document the importance of distinguishing between age and cohort effects. For example, there may be an effect on earnings of being 30 years old, or there may be an effect of having experienced the Great Depression. Data obtained at a single point in time (e.g., 1960) cannot distinguish the two effects; repeated surveys of the same people over time can sort out these cohort and age effects (see, e.g., Heckman and Robb, 1982; Weiss and Lillard, 1978).

Hause (1980) studied the time series properties of earnings in the context of the on-the-job training hypothesis. His results confirm the results of others who use cross-sectional, artificial cohorts, or less detailed panel data--namely, that there is a strong negative relationship between initial earnings and future wage growth. Controlling for unobservables and using longitudinal data, Hause finds that, among individuals of identical levels of schooling, those who have relatively low wages early in their careers are more likely to have relatively higher wages later in their careers. The statistical procedures used by Hause produce evidence of the pure life-cycle effects on the distribution of earnings emphasized by Mincer (1974).

Hours of Work

Over the last 20 years the analysis of labor supply has received substantially more attention than the study of labor demand. Most problems associated with low labor incomes have been analyzed in a context that emphasizes labor supply rather than the market interaction of both supply and demand forces. The economic analysis of the negative income tax experiments, for example, relies almost exclusively on models of the problem of single-period labor supply, which treat the pattern of available wage rate/income tax combinations as parametrically given.

The pioneering work on modern labor supply is Mincer's model of female labor force participation (Mincer, 1962a) and its extensions (Mincer, 1963). It is a life-cycle model of labor supply in which a woman, using a permanent wage measuring average earning power over her lifetime, chooses the fraction of her lifetime that she will work. By assuming that work time is perfectly substitutable across ages, Mincer converts the total lifetime labor supply function into a probability of observing a woman in the labor force at a particular time. Because of the assumption of perfect substitutability of time from one period to another, the age-specific female labor force participation equation depends only on the average lifetime wage.

Subsequent analysis (Bowen and Finegan, 1969; Cain, 1966; Kosters, 1966) generalizes certain aspects of Mincer's framework while introducing a restriction that is not in the spirit of his analysis. For example, Kosters uses a more rigorous demand system approach in specifying the labor supply behavior of both men and women in the context of a model of joint family decisions. In addition to estimating the usual substitution effect, Kosters attempts to measure the effect of taxes on labor supply directly. This work initiated a substantial empirical literature in which the labor supply effects of public policies were studied using a one-period model, in which the period is usually a single year in the working life. The vast literature on the evaluation of the income maintenance experiments uses this framework almost exclusively. The life-cycle focus of the original Mincer analysis disappeared. Mincer simplified a lifetime decision problem until it could be expressed in terms of a well-specified single-period decision problem; many of the subsequent studies presented single-period models that were no longer consistent with optimal lifetime behavior.

The one-period labor supply model has its theoretical origins in the work of Robbins (1930) and Lewis (1956). Even though this model is not in general consistent with a life-cycle labor supply model, the development of empirical models in the context of this single-period decision problem produced the first substantive discussion of the difference between estimating a labor-force participation model and an hours-of-work model. The analyses of Lewis (1969), Ben-Porath (1973), and Heckman (1974) demonstrate strong theoretical reasons why the effects of wage and nonlabor income on the labor force participation decision are different from the effects of these same variables on the hours-of-work decision. In addition, Heckman's (1974) work offered a statistical solution to the problem of selection bias--the problem that those who work and those who do not work are not necessarily random samples drawn from the same population. This work has been extended by Cogan (1981) to accommodate fixed costs of work (e.g., the fact that transportation costs to and from work affect decisions about hours of work). Recent papers on labor supply document the empirical importance of this problem of "selectivity bias" in estimating labor supply equations (see, e.g., the essays in Smith, 1980).

The problem of selection bias may arise whenever membership in a sample is generated by the choices of the individuals in the sample. It may also arise if sample membership is generated as a result of rules produced by administrators of social programs that the analyst is seeking to evaluate. In general, techniques that assume random sampling will not generate statistically desirable estimates when they are applied to such choice-censored samples.

Recent work on conventional labor supply models has adopted a multiperiod decision framework (see Heckman and MaCurdy, 1980; MaCurdy, 1978, 1981). The original Mincer insight that family labor supply behavior (particularly the labor supply of married women) is fundamentally a life-cycle problem has been salvaged. In recent studies the effects of past and future wage rates on current labor supply decisions have received their deserved attention. These studies build on earlier models of Ghez and Becker (1975), Heckman (1976), and Lucas and Rapping (1970) by considering the appropriate statistical framework when time is not perfectly substitutable across periods. However, the new dynamic models remain only an intermediate stage in the development of a general dynamic model of labor supply. Wage rates remain the only link between the worker and the labor market. Even though the full set of lifetime wage rates is used, no other characteristics of jobs or employers enter into the decision framework, and the possibility of long-term labor contracts is not considered. Life-cycle models have not addressed the problems of constraints on labor market hours, overtime, multiple job shifts, and the hedonic pricing of worker skills. All of the cited dynamic models assume that workers can borrow and lend freely over their lifetimes. With the exception of MaCurdy's model, they all assume an environment of perfect certainty. The effect of skill acquisition has not been fully integrated into the models, although some studies do allow for human capital investments in dynamic models of labor supply (Heckman, 1976; Ryder et al., 1976).

More Explicit Models of the Demand Side of the Labor Market

The labor supply research discussed above makes the fundamental assumption that a worker can sell all the hours he or she wishes at a given wage. A growing literature is challenging that assertion. Lewis (1969) suggests a model of employer determination of employee hours of work. Rosen (1974) formalized the market structure implicit in the "hedonic" approach. The market interaction of worker preferences and employer preferences for wage-hour packages produces an equilibrium locus of wage-hour packages in which the average hourly wage depends on the number of hours supplied.

In an attempt to bring demand factors into the analysis of the determinants of earnings, Tinbergen (1956, 1975) integrates the demand side of the labor market with the supply side. The central idea in his models is a specification of the relationship between worker skills and job tasks exploiting the comparative advantage of skill groups at each task (see also Rosen, 1978; Sattinger, 1975, 1980). The hidden demand variables such as capital, firm size, risk, etc., that stand behind the human capital regressions are made very explicit in this type of model.

Tinbergen (1975) examines the effect of supply and demand on the market for college-educated workers, demonstrating that there is considerable responsiveness of wage rates to quantities of labor demanded. He uses this model to explain how declining skill differentials arise. As quantities of skilled labor increase, the relative premium going to skilled labor diminishes. Tinbergen offers some empirical evidence on this issue and provides an interesting application of his models by investigating the effect of taxes on the distribution of income.

The study of the determinants of earnings is in its infancy. While the demand side of the labor market appears only in the most implicit form in most human capital models, Tinbergen's general equilibrium models of the labor market determination of earnings restore demand to the analysis of the determination of labor market earnings. This type of structural earnings function will prove useful in the analysis of labor market policy.

Turnover and Unemployment

The analysis of job search, labor market turnover, and job mobility has only recently become the focus of systematic empirical investigation. Sant (1977), Kiefer and Neumann (1981), and Flinn and Heckman (1982) have investigated the effect of dispersion in market wage offers on the search activities of individuals. A second area of recent increased activity is the relationship between job tenure and firm turnover. Imperfect information and the sorting and matching of workers play a critical role in explaining labor force dynamics (see especially Jovanovic, 1979a; MacDonald, 1980; and Mincer and Jovanovic, 1981). Just as imperfect information about the wage a worker can get at a given firm can account for unemployment, so the uncertainty about the potential gains from a match may cause a worker to work for an employer for a while to try to determine whether there is a good match. There has been very little empirical investigation of search and turnover using structural economic models. Little consensus exists on what unemployment is all about. The empirical search-theoretic models assume that search accounts for unemployment. The possibility of involuntary unemployment is typically not allowed.

A serious untested proposition of labor market behavior, which divides the neo-Keynesians (who assume pervasive involuntary unemployment) from the neoclassical economists and the more institutional from the less institutional labor economists, is the issue of whether the labor market is in equilibrium. Recent work by Ashenfelter (1980), Ashenfelter and Ham (1979), and Dickinson (1980) considers labor markets in which there may be some possibility of disequilibrium. The general view of the labor market held in these models is a neo-Keynesian view, espoused in the work of Barro and Grossman (1971). Individuals would like to select a given number of hours of work given their wage rates, but because of the distribution of demand in the labor market at large, individual offers of labor supply may not be purchased.

At issue is whether a worker's desired labor supply should be measured as unemployed hours plus actual hours worked. For example, Masters and Garfinkle (1977) have done this assuming that individuals are "involuntarily" thrown off their labor supply curves. Recent work by Chari (1980) and Grossman and Hart (1981) develops a labor supply theory that encompasses the idea that workers and firms may in fact have provisions for workers to spend part of the time less than fully employed and that these workers may be involuntarily unemployed (i.e., their incremental contribution to market output may exceed the value of their nonmarket time). These new models provide a neoclassical foundation for involuntary unemployment that, however, does not justify the sort of policy intervention, suggested by Keynesian models of involuntary unemployment, that implies that government policy may expand total output by employing workers until their incremental output equals the incremental value of their nonmarket time.

The disequilibrium view of the labor market has not been adequately tested. Tests of the neo-Keynesian models that have been pursued to date (e.g., Ashenfelter, 1980) suffer from serious identification problems (see Heckman et al., 1981). The issue of equilibrium versus disequilibrium in the labor market fundamentally affects how one interprets the estimated labor supply functions and how policies should be designed to affect the labor market, yet no convincing test between the two models has been devised, much less implemented.

The Emerging View of the Labor Market

The traditional view of the labor market, taught in standard economics texts and articulated in Hicks's Theory of Wages in 1932, is still the mainstay of a considerable body of work in labor economics. Hicks developed a model of the labor market that specifies a demand curve relating the wage rate paid to the total quantity of labor demanded by a firm or industry and a supply curve tracing a relationship between the wage and the total quantity of labor supplied. The intersection of the demand curve and the supply curve produces market equilibrium. This theoretical construct has proven quite useful for analyzing long-run problems; however, the empirical support for this very simple view of the labor market has not been strong, especially when it is applied to short-run issues. Empirical tests of models formulated from the perspective of homogeneous labor have yielded mixed results at best; reconciliation of cross-sectional and time series results have not yet been successful. Hicks's simple model of homogeneous labor is unable to explain a variety of important labor market phenomena, including job turnover, skill acquisition, labor hoarding, retirement, unemployment, and so forth. The analysis of these issues requires substantial modification of the theory, as Hicks himself notes (see Hicks, 1966, ch. V).

The appealing parallels between the analysis of labor services and of other marketed goods that result from thinking of labor services as a homogeneous good breaks down in application. The relevant characteristics of a bushel of wheat sold in a grain market are few and are relatively easily determined. The relevant characteristics of a unit of labor services sold in the labor market are many and are not at all easily determined--skills, attitudes, dependability, personality, flexibility, health, and so forth--and cannot be easily ascertained. Similarly, how a bushel of wheat is used after it is sold is in general of no concern to the seller, but not so with the sale of labor services. The importance of the conditions of the use of labor services, the environment in which they are used, and the broad package of remuneration in exchange for them are not typical of the sales of many goods (but see Carlton, 1979). Information about both sides of the market--about the labor services and about the environment and remuneration offered in exchange--is costly to obtain. This is true in part because neither the product sold (the labor service) nor the price paid (the package of employment conditions and remuneration) is a standard homogeneous entity. On both sides of the market a number of unique features exist because of the nonhomogeneous nature of the product and its price.

This view of the labor market can be found in a number of publications by economists of very different political persuasion. Without misconstruing many essential ideas by authors who range from Marxist labor economists, such as Bowles and Gintis, to dual labor market economists, such as Doeringer and Piore, to more neoclassical economists, such as Rosen, Starret, and Stiglitz, a common view of the labor market is beginning to emerge.

A key ingredient in much recent labor theory is the idea that workers differ in characteristics about which information is scarce. Jobs in turn differ in many important characteristics about which information is scarce as well. The heterogeneity of workers and the scarcity of information about workers imply that a different analysis of labor services is required than the analysis used to analyze a relatively homogeneous product like wheat. The issues of measuring product quality, monitoring and policing the worker's effort, providing the worker with appropriate incentives, and structuring the workplace to make the output of the individual greater are central to contemporary research in the economics of the labor market. Heterogeneity in working conditions and the package of remuneration offered by firms means that one has to treat the price of labor services in a different way than one treats the price of a homogeneous good like wheat. Here the issues of the training potential of the job, the wage-growth potential and the occupational growth potential of the job, the environmental context of the job (the quality of the air, the temperature, the noise), the nature of the job (its routinization, its physical demands, its pleasantness), and the remuneration package of fringe benefits, hourly wage, flexibility of hours, and so on are all components of the price paid or the offer made by the employer. Models and policies based on a unidimensional wage rate lack cogency in application to the labor market.

Signals

A recent paper by a radical economist, Gintis (1976), offers a coherent statement about the labor market that is not inconsistent with neoclassical analysis. Gintis's paper is representative of several recent contributions by authors in different segments of the economics literature who have independently articulated models that focus on the uniqueness of labor as a factor of production and stress that labor is not a homogeneous good. These models stress that there is imperfect information about labor and that labor market signals such as demographic characteristics convey information about average individuals in groups. (For development of these points in the theoretical industrial organization literature, see Spence, 1973.)

These ideas permeate modern labor economics and suggest why there are persistent differences in labor market outcomes of individuals of different demographic status. In an ideal world, in which information was free, signals would not be related to earnings except through their correlation with true productivity. In a world in which information about true productivity is scarce, these same signals may have considerable value as a measure of true, but unobserved, productivity.

Employer reliance on signals may, however, serve as an impediment to those individuals who are more productive than the average worker in the demographic group to which they belong, resulting in statistical discrimination. A very able person classified in a group with lower than average productivity may be treated as someone with lower productivity when, in fact, the opposite is true. The fact that information is imperfect and costly is one reason why the "true productivity" of the individual departs from the wage. It is difficult to assess the true productivity of the individual; individuals with a variety of productive attributes may not be able to reveal these attributes fully at any reasonable cost. This sort of imperfect information argument has attracted a great deal of attention in economic theory and is potentially relevant to the empirical analysis of the labor market. It may explain the persistence of discrimination (see, e.g., Spence, 1973; Spero and Harris, 1968; and Stiglitz, 1975) and indicates the potential value of information in the labor market.

The signaling hypothesis has been used to rationalize the measured effect of schooling on earnings. This model has been viewed as an alternative to the human capital model. Empirical studies by Layard and Psacharopoulos (1974), Taubman and Wales (1973), Wise (1975a), and Wolpin (1977) do not present firm evidence on the relative importance of the two hypotheses (see Riley, 1979). In fact it is not known whether any empirical test based on market transactions can distinguish the two hypotheses (short of access to ideal data about "true productivity").

Incentive Monitoring

A second aspect of costly information is the problem that an individual employer has in determining whether a worker is performing at the appropriate level. This is the incentive-monitoring problem (or principal-agent problem) that arises because it is costly--sometimes impossible--for the employer to measure the productivity of workers. The issue may arise because individuals are working in groups and it is possible to measure only the productivity of the group, or because the person who is most interested in measuring the productivity of the worker is far removed from the worker when he or she actually performs the task (or because it is very costly to monitor the performance of the task). In a world of complete information, incentive schemes would be irrelevant; when information about performance in the short run is very costly to acquire, such schemes develop. In most such schemes a worker's wage need not equal the value of the marginal product measured at a point in time.

One reason a worker's wage may not necessarily equal the value of his or her marginal product is that it is often advantageous to the firm to postpone some of the worker's payment, in order to police the worker for malfeasance or to encourage the worker to remain with the firm. The worker's performance, which in a standard neoclassical model is never considered, may improve if a deferred compensation scheme is used. If an employee is paid wages higher than productivity after a period in which he or she is paid wages below productivity, this pattern discourages malfeasance, for if malfeasance is detected before the high wage period the worker's wages can be confiscated to compensate the firm for the malfeasance, so the benefits of malfeasance to the worker will be diminished. Models of this kind can be found in Becker and Stigler (1974) and Gintis (1976) and applied to mandatory retirement in Lazear (1979b). An example of such deferred compensation is nonvested pensions.

In a world of imperfect information a worker may be directed by the employer to perform one of a set of tasks in any period. The allocation of workers to specific tasks may not be strictly determined by a market mechanism. These ideas are developed in theories of the firm by Simon (1957) and Alchian and Demsetz (1972) and are also found in Gintis (1976). Work by Becker and Stigler (1974) and others emphasizes the crucial point that there are many methods available to the firm to motivate their employees. Nonmarket transactions play a fundamental role in directing and in utilizing labor services in these analyses.

Matching

Because of the uniqueness of worker-firm matches there are incentives for firms to reward and employ workers in ways that they would not in the traditional neoclassical model. Consider firm-specific human capital, a concept that appeared in the human capital literature but that is clearly enunciated in various other literatures under different names. The idea is simple but important: A worker often acquires a skill of primary value only to the current employer. Information about the suitability of the worker for a particular job (or firm) is another example of job- (or firm-) specific capital. The existence of such private information is a basis for the sort of promotion and incentive schemes described in the internal labor market literature and provides an incentive for firms to hire or promote from within. By promoting from within a firm has access to information about the characteristics of workers that could not easily be obtained through simple hiring procedures. A firm may have an incentive to maintain an internal work force and put its employees through a promotion ladder in a hierarchy. This serves not only to train employees in the way that is most suitable for the firm, but also to sort or classify workers in a process that is intrinsically time-consuming but vital in securing the best worker-firm matches. Firm-specific capital or matching capital has received a great deal of attention and is a subject of ongoing theoretical research. From the original pioneering work of Becker (1975) to the more recent work of Jovanovic (1979b), Johnson (1978), and the extensions reported in Miller (1981), the idea of the matching of a worker to a job has become central.

Firm-worker match theories capture many of the essential consequences of employee and employer heterogeneity. A potential employee looks at conditions of employment that go beyond the simple descriptions of pecuniary compensation to incorporate nonpecuniary work conditions. Firms and workers enter into contracts. These contracts are sometimes made explicit, especially in sectors with unions. Contracts are made implicitly in all types of occupations. The employment relationship is not characterized by a simple quantity-for-price transaction, but rather is characterized by commitment of the worker to the firm and the firm to the worker. This intricate relationship is distinct from the degree of formality involved in the employment agreement. Promotion hierarchies emerge as a way of sorting, sifting, and training workers, thereby identifying highly productive workers, supervisors, and managers. The system tends to economize on costly information.

Coase (1937) stressed that a possibly more efficient way of organizing resources would be for workers to delegate some authority to managers. Instead of all functions within a firm being determined by a price mechanism, many firms might be managed simply by agreements of workers to be directed to certain tasks at the discretion of the firm. The setting of the boundaries concerning which tasks each worker will perform and what rules will govern that performance are still very much a subject of active discussion in the theoretical and empirical literature (e.g., Rosen, 1978, 1981; Sattinger, 1975, 1980).

Applications of the Emerging View

The view of the labor market and of the employee-employer match outlined above is, we suggest, an emerging view shared by many analysts. The essential features of this view are that labor markets are characterized by costly information, involve a heterogeneous commodity exchanged for a multidimensional package of remuneration and job conditions, and are governed by relatively complex, often implicit, long-term contracts. This perception of the labor market has important implications for the way in which many market phenomena are viewed. We suggest a few of these implications briefly.

Minimum Wages Suppose we take the traditional view that implies that firms hire homogeneous labor. Assume that firms experience an increase in the effective minimum wage. Under the traditional view, firms would be expected to employ fewer workers in response to the rise in the minimum wage (Stigler, 1946). In the newer model, there might still be a disemployment effect, at least in principle, but firms would have a richer menu of alternatives to consider. A firm that is providing some type of training might instead adjust the amount of training or adjust some other nonpecuniary attributes of the job. The firm might raise the wage and reduce the amount of training or raise the wage and charge explicitly for the training (see Barzel, 1976; Hashimoto, 1980; Mincer and Leighton, 1979). If the firm has acquired information about the worker and has invested in the worker's productivity with this firm, these information and training costs might attenuate the disemployment effects, at least in the short run. The employer might change the nature of the implicit contract with the employee rather than sever their match.

Nonpecuniary Compensation Instead of the worker's wage being the only dimension of compensation, the package of remuneration includes nonpecuniary benefits as well as disadvantageous job attributes such as health hazards. The standard theory of equalizing differences (Reder, 1962; Smith, 1776) suggests that workers who opt for compensation in nonpecuniary forms will necessarily forego wages. Therefore, a society in which more productive workers take relatively more of their compensation in nonpecuniary forms will exhibit less measured income equality than a society in which those more productive workers take their wages in relatively more pecuniary forms. Yet the distribution of earning potential could be the same in both societies.

Empirically, Duncan (1977) examines the effect on income of fringe benefits, health and safety, control of overtime hours, employment stability, and job autonomy. Similarly, Lucas (1972) and Brown (1980) look at the relationship between earnings and nonpecuniary job characteristics such as the requirements of physical strength and working conditions. Results of these studies are not consistent. Exposure to various types of risks on the job has been studied by Thaler and Rosen (1975). They quantified the extent to which firms may have to pay higher wages to compensate for job risks (see also Viscusi, 1978).

This work is relevant for policy studies because it suggests a much wider range of reaction by firms to policy interventions in the labor market. Suppose a firm encounters an increase in wage rates for some external reason. Firms can respond to that labor scarcity by varying a number of nonwage dimensions of employment conditions. Firms can change pension plans or the risk characteristics of the job or a variety of other factors. There are many competitive responses that do not alter the observed wage rate. Some of these responses may be more attractive to the firm than increasing the money wage. That is, the marginal cost of adjusting work conditions may be less than the marginal cost of increasing wage rates for a comparable increment in worker well-being. Similarly, a firm planning a permanent expansion might raise current wages to attract new workers, but the firm could instead raise pension levels or some other deferred compensation if its ultimate goal were to attract more workers over the longer term in order to save costs of turnover. These competitively viable responses indicate why the empirical assessment of wage employment dynamics is made very difficult without measurement of nonwage job characteristics.

Layoffs Other critical differences distinguish labor markets from other markets. In the conventional neoclassical model layoffs simply do not exist. A worker is either productive with the firm or is not productive with the firm. But if the firm has invested considerable information in finding out about the quality of the worker and in training the worker, it may not be optimal for a firm to sever all connections with the employee if it experiences an unfavorable demand shock. It may want to keep the worker and it may have to compensate the worker for the reduced demand. The deferred schemes common in the auto industry, in which unemployment insurance funds are maintained by the union but part of initial financing comes from the firm, are one example. The unemployment reserve fund keeps the workers attached to the firm during the layoff period.

Unions In a recent analysis of unionism by Freeman and Medoff (1978), the union is not viewed, as in the neoclassical model, as an agent of raising the wage of labor and thus causing some inefficient utilization of resources in the economy, but rather as a productive means of guaranteeing labor contracts, providing insurance, and revealing employee preferences (see also Stafford and Duncan, 1980). There is some very controversial empirical evidence in Brown and Medoff (1978) that unionism may raise the productivity of workers (controlling for movement up a firm's demand curve that results from higher wages). This view is not consistent with the traditional neoclassical view but deserves serious consideration as a coherent way of explaining some dimensions of trade union behavior. If true, it would certainly have implications for policies encouraging or discouraging union behavior.

The contrast with the conventional view does not contradict the evidence of Lewis (1963) or Rees (1962) on the observable effects of unions on wages. Instead, the questions require isolation of more aspects of the union-firm agreement in order to distinguish monopolization of labor from productivity gains that could result from information transmission and reduced turnover.

Hierarchies Consider the implications of the emerging view of labor markets for wage determination. Since job turnover is costly for both workers and firms, the effect of the supply side of the market in particular is predicted to be less immediate than would be posited in the traditional view of the labor market. Models of the internal labor market explicitly address these issues. These models were developed to explain the existence of job hierarchies within firms. Firms (especially large ones) tend to promote exclusively from within, only rarely going outside the firm. A number of labor market studies document this practice.

In this model firms are not totally isolated from the outside market. But the process by which firms are affected by changes in external labor market conditions is much slower than the traditional theory assumes. Firms use training programs and promotion hierarchies for workers to solve information problems, provide incentives against malfeasance, and to divide the gains from finding a productive match. Views of the labor market that suggest that employment of workers can be increased simply by giving them a particular skill in a training program are inadequate. They are based on a view of the firm that ignores how the promotion process works.

Research on this subject is under way in sociology in the work of Sorensen (1977) and Stewman and Konda (1982) and in economics in the work of Doeringer and Piori (1971) and Wise (1975b). A number of studies of hierarchies within firms have appeared, including work by Rosen (1978), Sattinger (1975, 1980), Stiglitz (1975), and Pettengill (1980). Work on hierarchies by Mirrlees (1976) poses the problem of choosing an optimal promotion system in a neoclassical production setting. Lazear and Rosen (1981) stress hierarchies in their analysis of high-level management positions.

Understanding the nature of how each job group functions is necessary to devise labor market models that give accurate measures of worker and firm responses to particular policies. The fact that research on this topic is being conducted by neoclassical economists, by radical economists, and by dual labor market economists is evidence that economists from a variety of ideological perspectives are converging on a common paradigm of the labor market, and that the phenomenon of internal labor markets and job hierarchies is both real and empirically important.

CONCLUSION

This paper describes the nature of the U.S. distribution of income and how it differs among several groups and over time. We have discussed the insights from recent theoretical literature on the determinants of labor earnings. We noted at the outset the linkages between the two parts of the paper are not especially close, reflecting the current gap between theory and fact. Indeed, one of the most important themes developed in Part 2 of this paper has been that the money wage, which constitutes a large portion of measured income, is not a full measure of the remuneration of employment. The lack of integration reflects the need for research to integrate more adequately the insights about labor markets into our understanding of the shape and causes of both earnings and income more broadly defined.

One of the important conclusions drawn in Part 1 is the crucial role played by demographic factors (e.g., by the age distribution and the distribution of family structure) in affecting the income distribution. Closely related is the point that the unit of analysis over which the distribution is observed (e.g., the individual, the earner, the family, or the household) greatly affects one's perception of the degree of inequality in income and of its changes over time.

In Part 2 we have presented the intellectual fruit of recent theoretical studies of labor supply, labor demand, and labor markets. We have stressed that despite apparent differences in ideology, a common view about the labor market is beginning to emerge.

REFERENCES

Aaron, H. J.
1978 Politics and the Professors: The Great Society in Perspective. Washington, D.C.: Brookings Institution.

Alchian, A. A., and H. Demsetz
1972 "Production, information costs and economic organization." American Economic Review 62 (5) (December):777-795.

Anderson, M.
1978 Welfare: The Political Economy of Welfare Reform in the United States. Stanford, Calif.: Hoover Institution Press.

Ashenfelter, O.
1980 "Unemployment as disequilibrium in a model of aggregate labor supply." Econometrica 48 (3) (April):547-564.

Ashenfelter, O., and J. Ham
1979 "Education, unemployment and earnings." Journal of Political Economy 87 (5) (Part 2, October):S99-S116.

Atkinson, A. B.
1970 "On the measurement of inequality." Journal of Economic Theory 2:244-263.

Barro, R. J., and H. I. Grossman
1971 "A general disequilibrium model of income and employment." American Economic Review 61 (1) (March):82-93.
Barzel, Y.
1976 "An alternative approach to the analysis of taxation." Journal of Political Economy 84(6):1177-1197.
Becker, G. S.
1960 "Underinvesting in college education?" American Economic Review 50 (2) (May):346-354.
Becker, G. S.
1975 Human Capital: A Theoretical and Empirical Analysis. 2nd Edition. New York: Columbia University Press for the National Bureau of Economic Research (first published in 1964).
Becker, G. S., and G. J. Stigler
1974 "Law enforcement, malfeasance, and compensation of enforcers." Journal of Legal Studies 3 (1) (January):1-18.
Ben-Porath, Y.
1967 "The production of human capital and the life cycle of earnings." Journal of Political Economy, 75 (4) (August):352-365.
1973 "Labor force participation rates and the supply of labor." Journal of Political Economy 81 (3) (May/June):697-704.
Benus, J.
1974 "Income Instability." In J. N. Morgan, ed., Five Thousand American Families. Vol. 1. Ann Arbor, Mich.: Institute for Social Research, University of Michigan.
Blaug, M.
1976 "Human capital theory: a slightly jaundiced survey." Journal of Economic Literature 14, (3) (September):827-855.
Blinder, A. S.
1980 "The level and distribution of economic well-being." In M. Feldstein, ed., The American Economy in Transition. Chicago: University of Chicago Press for the National Bureau of Economic Research.
Bowen, W. G., and T. A. Finegan
1969 The Economics of Labor Force Participation. Princeton, N.J.: Princeton University Press.

Brown, C.
1980 "Equalizing differences in the labor market." Quarterly Journal of Economics 94 (1) (February):113-134.

Brown, C., and J. Medoff
1978 "Trade unions in the production process." Journal of Political Economy 86 (3) (June): 355-378.

Cain, G.
1966 Labor Force Participation of Married Women. Chicago: University of Chicago Press.

Carlton, D. W.
1979 "Contracts, price rigidity, and market equilibrium." Journal of Political Economy 87 (5) (Part 1, October):1034-1062.

Chari, V. V.
1972 Implicit Contracts and Involuntary Unemployment. Northwestern University Discussion Paper #459. Center for Mathematical Study in Economics and Management Science, August.

Chiswick, B. R., and J. Mincer
1972 "Time series changes in personal income inequality." Journal of Political Economy 80 (3) (Part 2, May/June):S34-S66.

Chiswick, B. R., and J. A. O'Neill
1977 Human Resources and Income Distribution. New York: W. W. Norton and Co.

Coase, R. H.
1937 "The nature of the firm." Economica 4 (16) (November):386-405.

Cogan, J.
1981 "Fixed costs and labor supply." Econometrica.

Congressional Budget Office
1977 Poverty Status of Families Under Alternative Definitions of Income. Background Paper #17 (revised). Washington, D.C., June.

Danziger, S., R. Haveman, and E. Smolensky
1977 "The measurement and trend of inequality: comment." American Economic Review 67 (3) (June):505-512.

1979 Income Transfer Programs in the United States: An Analysis of Their Structure and Impacts. Paper presented to the Joint Economic Committee of the United States, Special Study on Economic Change. May.

Danziger, S., R. Havemen, and R. Plotnick
1981 "How income transfers affect work, savings and the income distribution." Journal of Economic Literature 19 (3) (September):975-1028.

Dickinson, J. G.
1980 "Parallel preference structures in labor supply and commodity demand: an adaptation of the Gorman polar form." Econometrica 48 (7) (November):1711-1726.

Doeringer, P. B., and M. Piore
1971 Internal Labor Markets and Manpower Analysis. Lexington, Mass.: D. C. Heath.

Duncan, G.
1977 "Labor market discrimination and non-pecuniary work rewards." In F. T. Juster, ed., Conference on the Distribution of Economic Well-Being. Cambridge, Mass.: Ballinger Publishing Company for the National Bureau of Economic Research.

Flinn, C., and J. Heckman
1982 "New methods for analyzing structural models of labor force dynamics." Journal of Econometrics, special issue on longitudinal analysis, January.

Freeman, R. B.
1978 "Time Series Evidence on Black Economic Progress: Shifts in Demand or Supply?" Discussion Paper #6321. Harvard Institute of Economic Research, Cambridge, Mass.

Freeman, R. B., and J. B. Medoff
1978 Substitution Between Production Labor and Other Inputs in Unionized and Nonunionized Manufacturing. Unpublished paper. Harvard University.

Fuchs, V. R.
1974 "Recent trends and long run prospects for female earnings." American Economic Review 64 (2) (May):236-242.

Ghez, G. R., and G. S. Becker
1975 The Allocation of Time and Goods Over the Life Cycle. New York: Columbia University Press for the National Bureau of Economic Research.

Gintis, H.
1976 "The nature of labor exchange and the theory of capitalist production." Journal of Radical Political Economics 8:36-54.

Goldberger, A.

1975 Statistical Inference in the Great IQ Debate. Discussion Paper #301-75. Institute for Research on Poverty, Madison, Wis.

Golladay, F. L., and R. H. Haveman

1977 The Economic Impacts of Tax-Transfer Policy. New York: Academic Press.

Griliches, Z.

1975 The Changing Economics of Education. Discussion Paper #426. Harvard Institute of Economic Research, July.

Grossman, S. J., and O. D. Hart

1981 "Implicit contracts, moral hazard, and unemployment." American Economic Review 71 (3) (May):301-307.

Haley, W.

1976 "Estimation of the earnings profile from optimal capital accumulation." Econometrica 44 (6) (November):1223-1238.

Hashimoto, M.

1980 Minimum Wage and Earnings Growth of Young Male Workers. Unpublished paper. University of Washington, April.

Hause, J.

1980 "On the fine structure of earnings and on-the-job training hypothesis." Econometrica 48 (4) (May):1013-1030.

Heckman, J. J.

1974 "Shadow prices, market wages, and labor supply." Econometrica 42 (4) (July):679-694.

1976 "A life cycle model of earnings, learning, and consumption." Journal of Political Economy 84 (4) (Part 2, August):S11-S44.

Heckman, J. J., and T. E. MaCurdy

1980 "A life-cycle model of female labor supply." The Review of Economic Studies 47 (1) (January):47-74.

Heckman, J. J., and R. Robb

1982 "The longitudinal analysis of earnings data." In J. Heckman and B. Singer, eds., Longitudinal Analysis of Labor Market Data. New York: Academic Press.

Heckman, J. J., M. Killingsworth, and T. E. MaCurdy

1981 "Empirical evidence on static labour supply models: a survey of recent developments." In Z. Hornstein and J. Grice, eds., The Economics of the Labor Market. London: Her Majesty's Stationery Office.

Hicks, J. R.

1966 The Theory of Wages. 2nd edition. New York: St. Martin's Press (first published in 1932).

Hill, C. R., and F. P. Stafford

1974 "Allocation of time to preschool children and educational opportunity." Journal of Human Resources 9 (3) (Summer):323-341.

Johnson, T.

1975 "Zealots and malingerers: results of firm specific human capital investments." Southern Economic Journal 41 (4) (April):613-626.

Johnson, W.

1978 "A theory of job shopping." Quarterly Journal of Economics 92 (2) (May):261-277.

Jovanovic, B.

1979a "Job matching and the theory of turnover." Journal of Political Economy 87 (5) (Part I, October):972-990.

1979b "Firm specific capital and turnover." Journal of Political Economy 87 (6) (December): 1246-1260.

Kakwani, N.

1980 Income Inequality and Poverty. New York: Oxford University Press.

Kiefer, N., and G. Neumann

1981 "Individual effects in a non-linear model: explicit treatment of heterogeneity in the empirical job search model." Econometrica 49 (4) (July):965-979.

Kiker, B. F.

1971 "The historical roots of the concept of human capital." In B. F. Kiker, ed., Investment in Human Capital. Columbia: University of South Carolina Press.

Killingsworth, M.

1982 "'Learning by doing' investment in training: a synthesis of two 'rival' models of the life-cycle." Review of Economic Literature.

Kohen, A. I., H. S. Parnes, and J. R. Shea

1975 "Income instability among young and middle-aged men." Pp. 151-207 in J. D. Smith, ed., The Personal Distribution of Income and Wealth. Vol. 39. New York: Columbia University Press for the National Bureau of Economic Research.

Kosters, M.

1966 Income and Substitution Effects in a Family

Labor Supply Model. P-3339. Rand Corporation, Santa Monica, Calif., December.

Layard, R., and G. Psacharopoulos
1974 "The screening hypothesis and returns to education." Journal of Political Economy 82 (5) (September/October):985-998.

Lazear, E. P.
1976 "Age, experience, and wage growth." American Economic Review 66 (4) (September):548-558.
1979a "Male-female wage differentials: has the government had any effect?" In C. B. Lloyd, E. S. Andrews, and C. L. Gilroy, eds., Women in the Labor Market. New York: Columbia University Press.
1979b "Why is there mandatory retirement?" Journal of Political Economy 87 (6) (December): 1261-1284.

Lazear, E. P., and R. T. Michael
1980 "Family size and the distribution of real per capita income." American Economic Review 70 (1) (March):91-107.

Lazear, E., and S. Rosen
1981 "Rank order tournaments as optimum labor contracts." Journal of Political Economy 89 (5) (October):841-864.

Leibowitz, A. S.
1974 "Home investments in children." Journal of Political Economy 82 (2) (Part 2, March/April): S111-S131.

Levy, F.
1976 "How big is the American underclass?" Working paper. Berkeley Graduate School of Public Policy, Berkeley, Calif., June.

Lewis, H. G.
1956 "Hours of work and hours of leisure." Annual Proceedings of the Industrial Relations Research Association:196-206.
1963 Unionism and Relative Wages in the United States: An Empirical Inquiry. Chicago: University of Chicago Press.
1969 "Interes del empleador en las horas de trabajo del empleado" (Employer interests in employee hours of work). Cuadernos de Economica 6 (18) (August):38-54.

Lillard, L. A.
1977 "Inequality: earnings versus human wealth." American Economic Review 67 (2) (March):42-53.

Lillard, L. A., and R. J. Willis
1978 "Dynamic aspects of earning mobility." Econometrica 46 (5) (September):985-1012.

Lucas, R. E. B.
1972 Working Conditions, Wage Rates, and Human Capital: A Hedonic Study. Unpublished Ph.D. dissertation, Massachusetts Institute of Technology.

Lucas, R. E., and L. Rapping
1970 "Real wages employment and inflation." In E. S. Phelps, ed., Microeconomic Foundations of Employment and Inflation Theory. New York: Norton and Co.

MacDonald, G. M.
1980 "Person-specific information in the labor market." Journal of Political Economy 88 (3) (June):578-597.

MaCurdy, T.
1978 Econometric Model of the Labor Supply in a Life Setting. Unpublished Ph.D. dissertation. University of Chicago.
1981 "An empirical model of labor supply in a life-cycle setting." Journal of Political Economy 89 (6) (December):1059-1085.

Mahoney, B. S.
1976 The Measure of Poverty. Washington, D.C.: Poverty Studies Task Force, Department of Health, Education, and Welfare.

Masters, S., and E. Garfinkle
1977 Estimating the Labor Supply Effects of Income Maintenance Alternatives. New York: Academic Press.

Michael, R. T.
1982 "Measuring non-monetary benefits of education: a survey." In W. W. McMahon and T. G. Geske, eds., Financing Education: Overcoming Inefficiency and Inequity. Urbana: University of Illinois Press.

Miller, R.
1981 "Matching and Turnover: Part I Theory." Economics Research Center/NORC Discussion Paper #81-11, September.

Mincer, J.
1958 "Investment in human capital and personal income distribution." Journal of Political Economy 66 (4) (July/August):281-302.

1962a "Labor force participation of married women." Pp. 63-105 in H. G. Lewis, ed., Aspects of Labor Economics. Universities-National Bureau Conference Series No. 14. Princeton, N.J.: Princeton University Press.

1962b "On-the-job training: costs, return, and some implications." Journal of Political Economy 70 (5) (Part 2, October):S50-S79.

1963 "Market prices, opportunity costs, and income effects." In C. Christ, ed., Measurement in Economics. Stanford, Calif.: Stanford University Press.

1970 "Distribution of labor incomes: a survey with special reference to the human capital approach." Journal of Economic Literature 8 (1) (March):1-26.

1974 Schooling, Experience, and Earnings. New York: National Bureau of Economic Research.

Mincer, J., and B. Jovanovic

1981 "Labor Mobility and Wages." Pp. 21-63 in S. Rosen, ed., Studies in Labor Markets. Chicago: University of Chicago Press for the National Bureau of Economic Research.

Mincer, J., and L. Leighton

1979 Effects of Minimum Wages and Human Capital Formation. National Bureau of Economic Research Working Paper #441.

Mirer, T.

1974 "Aspects of the variability of family Income." In G. J. Duncan and J. N. Morgan, eds., Five Thousand American Families--Patterns of Economic Progress. Volume IV. Ann Arbor, Mich.: Institute for Social Research, University of Michigan.

Mirrlees, J. A.

1976 "The optimal structure of incentives and authority within an organization." Bell Journal of Economics 7 (1) (Spring):105-131.

Muellbauer, J.

1977 "Testing the Barten model of household consumption effects and the cost of children." Economic Journal 87 (347) (September):460-487.

Paglin, M.

1975 "The measurement and trend of inequality: a basic revision." American Economic Review 65 (4) (September):598-609.

1977 Transfers in Kind: Their Impact on Poverty, 1959-1975. Unpublished paper. Hoover Institution. Stanford, Calif., October.

Pechman, J. A., and B. A. Okner

1974 Who Bears the Tax Burden? Washington, D.C.: Brookings Institution.

Pettengill, J. S.

1980 Labor Unions and the Inequality of Earned Income. Amsterdam: North Holland Publishing Company.

Plotnick, R. D., and F. Skidmore

1975 Progress Against Poverty: A Review of the 1964-1974 Decade. New York: Academic Press for the Poverty Policy Analysis Series, Institute for Research on Poverty.

Pyatt, G.

1976 "On the interpretation and disaggregation of Gini coefficients." Economic Journal 86 (342) (June):243-255.

Reder, M. W.

1962 "Wage differentials: theory and measurement." In H. G. Lewis, ed., Aspects of Labor Economics. Universities-National Bureau, Conference Series #14. Princeton, N.J.: Princeton University Press for National Bureau of Economic Research.

Rees, A.

1962 The Economics of Trade Unions. Chicago: University of Chicago Press.

Reynolds, M., and E. Smolensky

1977 Public Expenditures, Taxes and the Distribution of Income: The United States, 1950, 1961, 1970. New York: Academic Press.

Riley, J. G.

1979 "Testing the Educational Screening Hypothesis." Journal of Political Economy 87 (5) (Part 2, October):S227-252.

Robbins, L.

1930 "On the elasticity of demand for income in terms of effort." Economica 10 (29) (June): 123-129.

Rosen, S.

1974 "Hedonic prices and implicit markets: product differentiation in pure competition." Journal of Political Economy 80 (1) (January/February):34-55.

1976 "A theory of life earnings." Journal of Political Economy 84 (4) (Part 2, August): S45-S67.

1977 "Human capital: a survey of empirical research." In R. Ehrenberg, ed., Research in Labor Economics. Volume I. Greenwich, Conn.: JAI Press.

1978 "Substitution and division of labor." Economica 45 (179) (August):235-250.

1981 "The economics of superstars." American Economic Review 71 (5):845-858.

Ryder, H., F. Stafford, and P. Stephan

1976 "Labor, leisure and training over the life cycle." International Economic Review 17 (3) (October):651-674.

Sant, D. T.

1977 "Reservation wage rules and learning behavior." The Review of Economics and Statistics 59 (1) (February):43-49.

Sattinger, M.

1975 "Comparative advantage and the distributions of earnings and abilities." Econometrica 43 (3) (May):455-468.

1980 Capital and the Distribution of Labor Earnings. Amsterdam: North Holland Publishing Company.

Schultz, T. P.

1975 "Long term changes in personal income distribution: the theoretical approaches, evidence and explanations." Pp. 147-163 in D. M. Levine and M. J. Bane, eds., The Inequality Controversy: Schooling and Distributive Justice. New York: Basic Books.

Schultz, T. W.

1961 "Investment in human capital." American Economic Review 51 (1) (March):1-17.

Simon, H. A.

1957 Administrative Behavior. New York: Free Press.

Smith, A.

1776 An Inquiry Into the Nature and Causes of the Wealth of Nations. 1976 reprint. Chicago: University of Chicago Press.

Smith, J. P., ed.

1980 Female Labor Supply: Theory and Estimation. Princeton, N.J.: Princeton University Press.

Smith, J., and F. Welch

1978 Race Differences in Earnings: A Survey and

New Evidence. Report R2295-NSF. Rand Corporation, Santa Monica, Calif., March.

Sorensen, A. B.
1977 "The structure of inequality and the process of attainment." American Sociological Review 42 (6) (December):965-978.

Spence, M.
1973 "Job market signaling." Quarterly Journal of Economics 87 (3) (August):355-374.

Spero, S. D., and A. L. Harris
1968 Black Worker. New York: Atheneum Publishers.

Stafford, F. P., and G. J. Duncan
1980 "Do union members receive compensating differentials?" American Economic Review 70 (3) (June):355-371.

Stewman, S., and S. Konda
1982 "Careers and organizational labor markets: a demographic model of organizational behavior." American Journal of Sociology.

Stigler, G. J.
1946 "The economics of minimum wage legislation." American Economic Review 36 (3) (June):358-365.

Stiglitz, J. E.
1975 "Incentives, risk, and information: notes toward a theory of hierarchy." Bell Journal of Economics 6 (2) (Autumn):552-579.

Taubman, P.
1976 "Earnings, education, genetics, and environment." Journal of Human Resources 11 (4) (Fall):447-461.

Taubman, P., and W. J. Wales
1973 "Higher education, mental ability, and screening." Journal of Political Economy 81 (1) (January/February):28-55.

Thaler, R., and S. Rosen
1975 "The value of saving a life: evidence from the Labor Market." In N. Terleckyj, ed., Household Production and Consumption. New York: Columbia University Press for the National Bureau of Economic Research.

Tinbergen, J.
1956 "On the theory of income distribution." Weltvirtschaftliches Archiv 77:155-173.
1975 Income Distribution: Analysis and Policies. Amsterdam: North Holland Publishing Co.

Treas, J., and R. Walther
1978 "Family structure and the distribution of

family income," Social Forces 56 (3) (March): 866-880.

Urban Systems Research and Engineering, Inc.
1976 The Measure of Poverty. Technical Paper III. U.S. Department of Health, Education, and Welfare.

U.S. Department of Commerce, Bureau of the Census
Annual Current Population Reports on Consumer Income. Series P-60.
1980 Statistical Abstract of the United States, 1980. December.

U.S. Department of Health, Education, and Welfare
1976 The Measure of Poverty: A Report to Congress as Mandated by the Education Amendments of 1974. Washington, D.C.: U.S. Government Printing Office.

Viscusi, W. K.
1978 "Labor market valuations of life and limb: empirical evidence and policy implications." Public Policy 26 (3) (Summer):359-386.

Weiss, Y., and L. A. Lillard
1978 "Experience vintage and time effects in the growth of earnings: American scientists 1960-1970." Journal of Political Economy 86 (3) (June):427-448.

Welch, F.
1975 "Human capital theory: education, discrimination, and life cycles." American Economic Review 65 (2) (May):63-73.

Willis, R., and S. Rosen
1979 "Education and self-selection." Journal of Political Economy 87 (5) (Part 2, October): S7-S36.

Wise, D. A.
1975a "Personal attributes, job performance, and probability of promotion." Econometrica 43 (5-6) (September/November):913-932.
1975b "Academic achievement and job performance." American Economic Review 65 (3) (June):350-366.

Wolpin, K. I.
1977 "Education and screening." American Economic Review 67 (5) (December):949-958.

Cultural Meaning Systems

Roy G. D'Andrade

BACKGROUND

Between 1955 and 1960 the human sciences changed in a radical way. Before the 1950s the dominant paradigm was behaviorism, with its assumption that most things about people--such as personality, culture, and language--could be understood as complexes of stimulus and response connections. During the 1950s this paradigm was confronted across a number of disciplines. In psychology, Jerome Bruner, George Miller, and others developed cognitive and information-processing views of action and learning. In linguistics, Chomsky showed that Bloomfield's behavioristic concept of grammar could not in principle account for the capacities of natural language grammars (Chomsky, 1957). And in anthropology, Geertz, Goodenough, Hall, Schneider, Wallace, and others presented the argument that culture does not consist of behaviors, or even patterns of behavior, but rather of shared information or knowledge encoded in systems of symbols.

While this revolution was influenced by Europeans such as Piaget, Saussure, and later, Levi-Strauss, the main force of the revolt came, I believe, from the intellectual wave of ideas accompanying the development of the modern computer. One might have felt convinced when reading Skinner that the scientific study of people does not need concepts involving unobservable mental processes, such as thinking and feeling. Such a conviction was hard to hold, however, when computer programs were developed that played chess and solved logic problems. If computers could have programs, why couldn't people?

The major difference between the behaviorist and cognitive paradigms concerns the role of internal representations. In the behaviorist tradition what a creature does is, in large part, controlled by various external conditions, such as the presence of conditioned and unconditioned stimuli and the number of hours of deprivation. In the cognitive paradigm what a creature does is, in large part, a function of the creature's internal representation of its environment. For many anthropologists, the emphasis of the cognitive paradigm on internal representations had a better fit to their intuitions about the nature of culture than behaviorist notions of stimulus control.

The conception of culture as knowledge and symbol rather than habit and behavior was rapidly assimilated into anthropology and the human sciences. Culture came to be seen as an information-holding system with functions similar to that of cellular DNA. For individual cells DNA provides the information needed for self-regulation and specialized growth. For humans, the instructions needed for coping with the environment and performing specialized roles is provided in learned information, which is symbolically encoded and culturally transmitted.

In considering the concept of culture from a cognitive perspective, this paper examines several current positions and related theoretical issues. First, the characterization of culture as a body of knowledge is discussed, and the criticisms of this position are reviewed with special reference to the part of constitutive rules in culture. Related to these issues, the treatment of cultural meaning systems as purely representational in character is criticized, and the argument is advanced that meaning systems have directive and evocative as well as representational functions. Problems with the current use of the term *symbol* are discussed, along with the difficulties involved in treating meaning systems as if they existed solely in external messages. The unnoticed development of a body of experimental techniques in the investigation of meaning systems by anthropologists and other social scientists is reviewed. Finally, the relationships between culture, social structure, personality, and experience are examined, and a definition of culture presented.

CULTURE AND CULTURALLY CONSTRUCTED THINGS

The initial cognitive formulations of culture focused on *knowledge*. "A society's culture consists of whatever it is one has to know or believe in order to operate in a manner acceptable to its members" (Goodenough, 1957). In

Goodenough's framework, knowledge typically consists of rules--rules by which one decides where to live, how kin are to be classified, how deference is to be expressed, etc. Thus, just as a computer operates by means of a program consisting of a set of rules that prescribe what actions are to be taken under various conditions, so the individual can be seen as operating by means of a cultural program. While Goodenough did not adopt the information-processing terminology and flow chart formats of computer science, such a vocabulary and set of formats were developed by others in the ethnoscience and cognitive anthropology tradition (e.g., Geoghegan, 1971).

The "culture as knowledge" formulation proved to have considerable potential for ethnographic investigation and theoretical analysis. Cultural knowledge about plants, animals, land use, navigation, etc., proved to be rich areas for ethnographic description. More theoretically, the idea that the complexity and heterogeneity of observed behavior could be accounted for by a small number of rules led to the development of formal and quasi-formal decision-making models capable of generating complex outputs from the interaction of a small number of external inputs and internal rules. Descriptively adequate and psychologically plausible models were developed to account for such things as kin term systems, patterns of residence, market choices, and legal fines. (For a review and critique of these models, see Quinn, 1975.)

While the conception of culture as consisting of the shared knowledge of individual minds marked a clear advance over earlier theories of culture, problems and attendant dissatisfactions quickly arose, becoming prominent by the 1970s. Three major problems became apparent: first, many things one would want to call cultural are not completely or even generally shared; second, culture consists of more than just knowledge; and third, it is not clear whether cultural systems are to be found "inside" or "outside" the minds of individuals. The last two of these issues were nicely caught in Geertz's example of a Beethoven quartet (1973:11-12):

> If . . . we take, say, a Beethoven quartet as an, admittedly rather special but for these purposes, nicely illustrative, sample of culture, no one would, I think, identify it with its score, with the skills and knowledge needed to play it, with the understanding of it possessed by its performers or auditors, nor . . . with a particular performance

> of it or with some mysterious entity transcending material existence. The "no one" is perhaps too strong here, for there are always incorrigibles. But that a Beethoven quartet is a temporally developed tonal structure, a coherent sequence of modeled sound--in a word, music--and not anybody's knowledge of or belief about anything, including how to play it, is a proposition to which most people are, upon reflection, likely to assent.

To continue the argument with a different example: marriage is part of American culture, but marriage is not the same thing as knowing how to marry people, or knowing how to get married, or understanding what it is to be married. Most Americans have an understanding of what banishment is, and how to banish someone (were they Richard II), yet these understandings do not make banishment a part of American culture.

If marriage is not the same thing as knowing about marriage, what is it? According to John Searle, marriage is a special kind of fact (1969:51-52):

> Any newspaper records facts of the following sorts: Mr. Smith married Miss Jones; the Dodgers beat the Giants three to two in eleven innings; Green was convicted of larceny; and Congress passed the Appropriations Bill. . . . There is no simple set of statements about physical or psychological properties of states of affairs to which the statements of facts such as these are reducible. A marriage ceremony, a baseball game, a trial, and a legislative action involve a variety of physical movements, states, and raw feels, but . . . the physical events and raw feels only count as parts of such events given certain other conditions and against a background of certain kinds of institutions. . . . It is only given the institution of marriage that certain forms of behavior constitute Mr. Smith's marrying Miss Jones. Similarly, it is only given the institution of baseball that certain movements by certain men constitute the Dodgers' beating the Cubs 3 to 2 in eleven innings.
>
> These "institutions" are systems of constitutive rules. Every institutional fact is underlain by a (system of) of rule(s) of the form "X counts as Y in context C."

Marriage is a part of American culture in that there is a constitutive system of rules that individuals know, which are intersubjectively shared and which are adhered to. Enactment of certain behaviors counts in certain contexts as "getting married," and once married, certain obligations and commitments are incurred. Marriage is a culturally created entity--*an entity created by the social agreement that something counts as that entity*. To agree that something will count as something else is more than simply knowing about it, although knowing about it is a necessary precondition. The *agreement* that something counts as something else involves the *adherence* of a group of people to a *constitutive rule* and to the entailments incurred by the application of the rule.

Probably every cultural category creates an entity, in the sense that what is understood to be "out there" is affected by the culturally based associations built into the category system. The English language cultural categories of *stone*, *tree*, and *hand* invoke a variety of shared connotations about these objects that add to whatever may be their reality as brute facts, but these cultural connotations do not manufacture the objects themselves from thin air. The cultural categories of *marriage*, *money*, and *theft*, on the other hand, are created solely by adherence to the constitutive rule systems that define them. Without these rule systems these objects would not exist.

Games make the most effective illustrations of constitutive rule systems, perhaps because the arbitrary nature of games makes the separation between the physical events of the game and what these events count as quite apparent. When a football player is declared "out of bounds," everyone understands that the physical fact of stepping over the line counts as being out of bounds only with respect to the game being played. However, everyday events like the theft of a sum of money are more likely to be treated as plain physical facts. To see theft as a culturally created entity one must be able to isolate the system of rules about what constitutes theft from the fact of physical removal and realize that physical removal of an object counts as theft only if certain conditions about intentions and ownership are satisfied.

A large number of the variables of social science refer to culturally created things. Family, property, deviance, prestige, race, and nationhood, for example, are all created by social agreement about what counts as what. The point is not an obvious one: Various anthropologists have had an uphill battle trying to convince the rest of the field that what is called kinship, for example, is created by a system of constitutive rules, not simply by facts of nature (Schneider, 1968).

Not all social science variables refer to culturally created things; some variables refer to objects and events that exist prior to and independent of their definition: for example, a person's age, the number of calories consumed during a meal, the number of chairs in a room, the pain someone felt, etc. Searle, following Anscombe (1958), calls the existence of such things "brute facts," in contrast to "institutional facts." Some social science variables, however, are not clearly either one or the other. For example, in many cases it is not clear whether the term social class refers to a set of cultural categories, which create the very thing they define, or to a culturally postulated entity, which exists independently of any cultural categories--or to certain aspects of both.

It is not just social scientists who are unclear about these matters. People often believe it is natural for women or fathers or Indians to act in certain ways. The problem is not whether these classes of persons have culturally constructed roles--most people agree that some part of what persons in these classes do is culturally learned role behavior--but rather which parts are roles created and which parts are natural expressions of character. It seems to be the case that people have a tendency to treat culturally created things as if they were natural things, perhaps because what is culturally created is often intricately intertwined with what occurs naturally and perhaps because it gives greater moral force to the idea that one should act in some certain way if it is thought that it is natural to act in that particular way. Thus, if one thinks of constitutive rules as culturally based "verdicts" about what counts as what, one can often find these verdicts behind what seems like naturalistic observation. For example, as Schneider (1968) has pointed out, the "observation" that kinship is made of flesh and blood contains the verdict that the physical facts of biological relatedness count as shared identity, which then entails the presumption that certain kinds of rights and duties will be assumed between "kin" as a matter of course.

While constitutive rules create entities out of "thin air," these entities are often embodied in physical tokens. For example, flags, capitals, and uniforms are treated as the embodiments of a nation state, paper bills are treated as embodiments of wealth, signatures are considered to represent personal commitment. Part of the extensive embodiment of constitutive entities seems to be a matter of practicality--it is easier to play chess with ivory pieces than to try to hold the game in one's mind, writing makes it possible to freeze talk in a timeless mode, and tokens like money and checks are a great convenience. Such embodiments as flags and uniforms also serve to create awe and respect, which, as Bentham pointed out, can be an advantage for those who rule.

Most constitutive rules are organized in a series of hierarchically linked systems. A memo, for example, involves a hierarchy of constitutive systems that link letters to sounds, sounds to words, words to sentences, and sentences to speech acts such as requests, commitments, etc. Constitutive rules are not only linked hierarchically but are also organized into elaborate systems, creating whole complexes of cultural entities. Thus the football complex creates touchdowns, quarterbacks, field goals, offsides, downs, etc. The family complex creates marriages, mothers, fathers, homes, joint property, relatives, in-laws, incest taboos, adultery, divorce, alimony, inheritance, breaking away, get-togethers, etc. Furthermore, constitutive systems tend to interpenetrate; for example, the family complex is intertwined with the property complex, the legal complex, and the religious complex. Schneider (1976) has termed these complexes galaxies and discussed how the complexes that include nationality, religion, locality, ethnicity, and family interpenetrate in American culture.

It is of some interest that Searle and Schneider should both have pressed the point that constitutive rules (in Schneider's terms culture as constituted) are to be distinguished from regulatory rules (Schneider's norms). According to Searle (1969:33-34):

> Regulative rules regulate antecedently or independently existing forms of behavior; for example, many rules of etiquette regulate inter-personal relationships which exist independently of the rules. But constitutive rules do not merely regulate, they create or define new forms of behavior. . . . Regulative rules characteristically take the form of or can be paraphrased as imperatives, e.g., "When cutting food, hold the knife in the right hand," or "Officers must wear ties at dinner." Some constitutive rules take quite a different form, e.g., "A checkmate is made when the king is attacked in such a way that no move will leave it unattacked. . . ."

According to Schneider (1976:202-203):

> Culture contrasts with norms in that norms are oriented to patterns for action, whereas culture constitutes a body of definitions, premises, statements, postulates, presumptions, propositions, and perceptions about the nature of the universe and mans place in it. Where norms tell the actor how to play the scene, culture tells the actor how the scene is set and what it all means. Where norms tell the actor how to behave in the presence of ghosts, gods, and human beings, culture tells the actors what ghosts, gods, and human beings are and what they are all about.

Basically, both Schneider and Searle see the distinction between the constitutive and the regulatory as a contrast between ideas that create realities and ideas that order or constrain action. The distinction is a necessary one if one wishes to analyze the relation between meanings and action, although each tends to be linked to the other, in that regulatory rules tend to be linked to the entities created by constitutive rules. For example, in a game of checkers, if a piece counts as a king this means it can move in either direction. In the world of property, if an object is sold, this means the seller no longer has certain rights over the object. Such entailments come as part of the very definition of the entity, so that what is being constructed is not just an object, event, or relationship, but is also a set of rules about what follows, given that something counts as that object, event, or relationship. Thus if war is declared in the United States, this declaration has a complex set of entailments concerning the powers of the President and the duties of citizens. It is a basic part of constitutive rule systems that the entities created have entailments to norms, and norms in turn entail action. Such entailments are not a matter of logic, but rather consist of the assumption that such linkages exist (Friedrich, 1977).

Just as most, if not all, constitutive entities entail certain norms of action, so most, if not all, norms are linked to certain constitutive rules. Wearing a tie, conducting an exorcism, or holding a knife in the right hand are linked to constitutive rules by which formality is defined and created, by which the notion of a spirit that can inhabit the body of a person is defined and created, or by which the notion of politeness is defined and created. Whole systems of norms, such as a kinship system or political system, are linked to whole systems of constitutive rules. This linkage is not a matter of logical necessity--very similar constitutive rules can be linked to quite different norms. Thus, for example, exactly which norms follow from the constitutive fact that two persons are of the same flesh and blood may vary quite widely in different subcultures in the United States (Schneider, 1968).

One consequence of constitutive rule systems is the enormous expansion of the behavioral repertoire of humans compared with the behavioral repertoires of other animals. For example, without the system of constitutive rules called football, the behaviors of scoring, blocking, passing, etc., would not exist. Without the constitutive systems of morality, etiquette, and efficiency, the behaviors of cheating, being rude, slacking off, etc., would not exist (Much and Shweder, 1978). Even the common and basic interpersonal acts of asserting, agreeing, requesting, and promising would not exist without the system of constitutive rules for speech acts (Austin, 1962; Searle, 1969, 1978; Vendler, 1972).

In talking about the creation of realities, it should be understood that what is being proposed is not the creation of realities of the type popularized by Carlos Castenada. Castenada appears to be proposing that there are alternative physical realities--as if under special conditions people can magically transform themselves into other people, fly through the air, affect others with their thoughts, etc. What is being proposed here is not that people can magically transform physical reality, but that people can create conventions, such as legality, nationality, marriage, etc., which are then taken account of as facts--something that exists.

A class of culturally created entities that I have been attempting to analyze involves the domain of success. This domain includes a number of elements referred to by such terms as accomplishment, recognition, prestige, self-satisfaction, goals, ability, hard work, competition, and the like. In American culture success is a personal characteristic of great importance to most people. Such daily events as the organization of daily effort, the evaluation of task performance, and the marking of accomplishment through self-announcement and the congratulations of others are closely attended to and much discussed.

A number of the elements of the world of success appear to be connected to each other through putative causal relations. Certain things are thought to lead to success, while other things are thought to result from success. Based on the initial data I have collected, it seems to be the case that Americans think that if one has ability, and if, because of competition or one's own strong drive one works hard at achieving high goals, one will reach an outstanding level of accomplishment. And when one reaches this level one will be recognized as a success, which brings prestige and self-satisfaction.

In success, the boundary line that divides a high from an ordinary level of accomplishment is not precisely specified. Often people do not know if they are really a success until some special award or position has been granted. This problem--deciding exactly what fits under the constitutive rule--appears to be endemic in social life. In a personal communication on this topic, Aaron Cicourel has pointed out:

> Searle's use of the term "constitutive rule" refers more to the general ideals or beliefs we share about a marriage ceremony, a baseball game, a trial or a legislative action than to the daily organized practices that produce marriages, baseball games, trials and legislative action. . . . John Rawls' distinction (1955) between a general rule or policy and a particular case said to fall under the rule can be instructive here. The constitutive and normative rules making up institutions do not provide instructions to members of a group on how to decide which daily life activities are constitutive. A policeman, for example, when making an arrest, is usually responding to a particular case as viewed under local contextual conditions and general personal conceptions of what is right and wrong. In order to justify his actions, the policeman must find a general rule or law statute to validate his actions with the particular case (Cicourel, 1968). There are many situations where this duality of knowledge about general rules and deciding that a particular case falls under one or more of them is crucially apparent. We are often confronted with situations in which our knowledge of constitutive rules becomes strained because of particular practices or cases that do not neatly fit any normal case. We tend to be comfortable with idealization

> in our everyday talk about institutions, but seldom examine the "normalization" required to use normative categories when we encounter discrepancies in practice. Adherence to a constitutive rule is a variable accomplishment. Interpretive procedures are needed for members to link constitutive rules with daily practices and vice-versa.

The duality Cicourel refers to contributes to making the ongoing process of social and cultural life a matter of at least occasional dispute and negotiation. One may be quite clear that X counts as Y, but it is often difficult to decide whether one is actually in the presence of a true X. To use a constitutive rule that X counts as Y requires the dual rule that X can be identified by the presence of features f1 . . . fn. And often some of the features f1 . . . fn are missing, ambiguous, or disputable, making it problematic whether or not something is an X, thereby making it problematic whether one is in the presence of a Y.

Indeed, some of the deepest social conflicts occur over the issue of the scope of a particular constitutive rule. Current debates about abortion and the rights of the fetus, the equality of women, the determination of comparable worth for different jobs, the two-mile ocean limit to national sovereignty, the age at which a person can appropriately engage in sexual activity, etc., attest to the importance of the determination of the scope of constitutive rules in social change. Debate about which constitutive rule is the right rule is rarely if ever decidable by means of logic or physical fact, although there are, I believe, empathy-based standards that can be used to decide some questions of this type.

THE FUNCTIONS OF MEANING SYSTEMS

In saying that success is a personal characteristic of great importance to most people, I mean to say more than that people think frequently about success. The argument I wish to make is that the meaning system involving the world of success, like most meaning systems, does more than represent facts and create entities. How one thinks of meaning depends on what one thinks meanings do. Meanings in general, and cultural meaning systems in particular, do at least four different things. Meanings represent the world, create cultural entities, direct one to do certain things, and evoke certain feelings. These four functions of meaning--the representational, the constructive, the directive, and the evocative--are differentially elaborated in particular cultural meaning systems but are always present to some degree in any system.

The current view, with some exceptions, treats meaning as having only representatational functions. From the representational point of view culture consists of knowledge and belief about the world, carried by true or false propositions composed of terms whose definition rests on potentially observable characteristics. This view has certain merits. First, cultural meaning systems generally have strong representational functions--with some exceptions, such as music, some of the arts, and ritual. Second, the representational function has great adaptational value--culture consists at least of knowledge about what is out there and what can be done with it, and this knowledge is carried through representation.

It seems clear, however, that most systems of meaning that are culturally acquired are not purely representational. In the discussion above concerning constitutive rules it was pointed out that cultural entities like marriage could not be created just from representational understandings. Besides understanding what counts as what, the creation of cultural things logically requires that people be bound to count X as Y and accept the entailments that follow. Most of us are bound to use words as they are normally understood, to accept paper money for our labor, to take responsibility for our kin--otherwise words would not be words, money would not be money, kin would not be kin, etc.

It needs to be stressed that learning a meaning system does not result in the learner's automatically and involuntarily following rules. Rather, various elements of the meaning system come to have a directive force, experienced by the person as needs or obligations to do something. For example, when the ordinary event of being asked a question occurs, one is not automatically impelled to answer, but the effect of normal socialization is that we experience a strong pressure to give an answer. Similarly, marriage in the United States involves a complex set of commitments that have directive force on individuals, commitments that are consciously made and experienced individually as powerful obligations (Quinn, 1980).

The assertion that people feel strong obligations and pressures as a result of socialization is not an unusual claim. Often, however, such obligations and pressures are treated by social theorists as if they were generated entirely by sanctions external to the individual. Spiro (1961:95-106) has pointed out that the antipsychological position of those who believe that conformity to cultural norms is due not to individual motivation but to external social sanctions in fact contains implicitly the psychological theory that people are, as individuals, motivated by exactly these sanctions.

A related theory proposes that the directives of cultural rules are based on the individual's generalized desire to conform to whatever it is that other people are observed to do or whatever it is other people say one should do, rather than on what people really want to do. While both external sanctions and conformity pressures certainly occur as means of social control and are probably necessary, the empirical evidence indicates that these kinds of extrinsic motivators are rarely the primary type of control in any society and are most prevalent in those historical periods marked by social anomie and mental pathology (Spiro, 1961:103).

More commonly or typically, the goals stipulated in the cultural meaning system are intrinsically rewarding; that is, through the process of socialization individuals come to find achieving culturally prescribed goals and following cultural directives to be motivationally satisfying and to find not achieving such goals or following such directives to be anxiety producing (Spiro, 1961:104-105). There appear to be two major intrinsic motivational systems involved with cultural meaning systems: the first is relatively direct personal reward; the second is reward because of attachment to a particular set of values. Typically the two are mixed: for example, in the cultural meaning system involving success, accomplishment may be rewarding both because it satisfies personal needs for recognition, achievement, security, etc., and because it represents the "good" self.

In general the directive functions of most cultural meaning systems are highly overdetermined: overdetermined in the sense that social sanctions, plus pressure for conformity, plus intrinsic direct reward, plus values, are all likely to act together to give a particular meaning system its directive force. For example, consider again the American meaning system of success. There are external sanctions involving money and employment, there are conformity pressures of many kinds, and there are the direct personal rewards and value satisfactions already mentioned. Perhaps what is surprising is that anyone can resist the directive force of such a system--that there are incorrigibles.

It may be objected that what I have been calling directive functions are more a part of personality and psychology than of culture and anthropology. After all, aren't things like goals and values part of an individual's personality, involving complex psychological processes like the formation of motives and the avoidance of anxiety? Doesn't the connection between symbol and motive come about only after elaborate socialization experiences, and aren't there many cultural symbols that are unrelated to motives? Isn't the inclusion of *directive* and *affective* functions in cultural systems of any sort a confusion of individual and cultural levels of analysis?

These objections are based on the assumption that for something to be truly cultural is must be acquired and performed without any significant involvement of psychological processes. There appears to be an implicit assumption in anthropology that anything that is known to involve complex psychological processes cannot also be cultural. Thus attitudes, needs, goals, defenses, etc., because they clearly involve complex psychological processes, are typically considered to be part of personality, not culture, no matter how shared or insitutionalized a particular attitude, need, goal, defense, etc., may be.

What is not appreciated is that most human behavior involves complex psychological processes. Take, for example, the formulation that culture is primarily knowledge. It is widely assumed that knowledge can be transmitted without involving psychological processes to any significant extent--someone tells someone else something, then the other person knows it: Simple communication through the transmission of information has occurred. George Lakoff has discussed this metaphor of transmission in detail and has indicated some of the confusions it engenders (Lakoff and Johnson, 1980; see also Reddy, 1976).

According to the transmission metaphor of communication, the speaker puts ideas into words (like objects put into boxes) and sends them via voice or letter to a hearer, who gets the ideas from the words (more or less as the speaker packed them). The transmission of objects in boxes requires no psychological processing--all that is needed is a physical system of moving boxes. But this is not true of ideas. For ideas to be communicated there must be a set of psychological mechanisms by which meanings are mapped into and out of physical signals--a procedure that is, so far as we now understand it, both complex and problematic.

There are a number of reasons why it is easy to overlook the psychological processing involved in the transmission of knowledge. First, once a person understands a language, it is easy to forget how much learning went into acquiring it. Second, it is easy to overlook the psychological processing that operates as information is transmitted, since the processes involved in understanding and producing speech are usually out of awareness and highly automatic. In general, there appears to be a sharp difference between the kind of psychological processes that are involved in cognition and perception and the kind of processes that are involved in motivation and feeling. When someone tells one something, it seems as if ideas come automatically with the words, while feelings and desires are experienced as aroused from some place within the mind, separate from the perceived symbol. But however these things seem, it is, I argue, the case that ideas, feelings, and intentions are all activated by symbols and are thus part of the meaning of symbols.

In general, there are a variety of lines of evidence that indicate that any human system of meaning is likely to involve affect. First, humans appear to have an affective response to almost any stimulus, no matter how decontexualized. Even a small patch of colored paint on a sheet of paper seems capable of arousing distinct and well-shared affective responses (D'Andrade and Egan, 1974). Second, some symbolic forms, such as poetry and music, clearly arouse strong and well-organized affective responses. Third, in ordinary speech there is a rich variety of expressive and evocative forms, such as thanks, apologies, condemnations, regrets, condolences, curses, congratulations, exclamations, cries, cheers, etc. All these kinds of evidence indicate that there is an emotional side to meaning. Often the evocative function blends with the directive function into a powerful good-happy-like approach versus a bad-fright/anger-dislike-hit/flee attitude, as Osgood's work with the semantic differential, in which these terms are commonly found together, demonstrates (Osgood et al., 1977).

One objection to the postulation that meanings have an affective function is the "affective recall thesis," which says that the affective responses of people when using symbols is due just to the natural or learned emotional reactions people have to the things referred to by the symbols. Thus, for example, it could be said that the exciting quality of the symbols that involve success are due to the excitement originally instigated by the events being referred to, and recalled by the use of symbols that refer to these events.

It seems very likely that the process of affective recall does occur. It seems unlikely, however, that this process alone can account for the kind and degree of feeling aroused by symbols. An often-mentioned counterexample to the affective recall argument illustrates that there can be different affective content to words that refer to the same physical object. For example, each of the terms _feces_, _shit_, _poo_, _stool_, _crap_, _excrement_, _turd_, etc., has a distinct kind of affective charge, although all these terms refer to the same thing. Each symbol appears to be a condensation of a number of affectively linked associations within a meaning system that cannot be explained on a simple experiential basis. Furthermore, it would be very difficult to imagine how the widespread personal agreement about the affective qualities of each of these terms could be arrived at through similarities in the physical experience of individuals.

In summary, the general position presented here is that meanings involve the total human psyche, not just the part of us that knows things. Every aspect of meaning systems requires a great deal of psychological processing and often considerable experiential priming. It takes years of learning for a child to acquire the representational functions of meaning systems. Representation occurs only because symbols activate complex psychological processes. In the same way, it takes years of learning for a child to acquire the constructive, directive, and evocative functions of meaning systems, and these functions, too, require complex psychological processes. The representational, constructive, directive, and evocative functions are each a consequence of the way the human brain is organized, a biological and psychological potentiality that is highly elaborated and stimulated by cultural meaning systems.

I have presented elsewhere the thesis that part of the tendency to play down the affective aspect of culture is based on the widely shared assumption that reason and emotion are basically in conflict and that emotion comes from the more animal and less advanced part of the human psyche (Shweder, 1981). I believe this thesis is wrong, that thinking and feeling are parallel processes that have evolved together because both are needed for any animal to attend to its needs in a highly intelligent way

(D'Andrade, 1981). Both processes tell us about how the world is. Sometimes both agree, and sometimes they disagree about what is the case and what should be done about it. The Socratic metaphor of reason and passion as two horses pulling a chariot seems much more accurate than the current metaphor of the war of reason and emotion. In any case, the assumption that reason and feeling are essentially and basically in conflict appears to be deeply ingrained in American and European culture, reinforcing the assumption that meaning is--or should be--entirely representational.

Others have expressed some of the views presented here at earlier times. In 1962, Clifford Geertz, in his essay on "The Growth of Culture and the Evolution of Mind," stated (p. 81):

> Not only ideas, but emotions too, are cultural artifacts in man. . . . The kind of cultural symbols that serve the intellective and affective sides of human mentality tend to differ--discursive language, experimental routines, mathematics, and so on, on one hand; myth, ritual and art on the other. But the contrast should not be drawn too sharply: mathematics has its affective uses, poetry its intellectual; and the difference in any case is only functional, not substantial.

When I first read this passage, I did not so much disagree as believe it to be irrelevant to the kind of problems I was working on. The affective, directive aspect of symbols, I thought, was to be found in religion and art--in the kinds of symbols Levy (1981) has called "marked symbols," in contrast to the commonsense world of "embedded symbols," such as kinship terminologies or classifications of illness. What I did not see was that my model of meaning led me to select for analysis that part of any system of symbols that was most representational and least affective or directive. Thus the componential analyses of kinship terminologies I was using could account beautifully for the way in which kin types are categorized but could not account for such simple aspects of meaning as understanding what is meant by the phrase, "Susan is a good mother." It was not until I came to appreciate that even (and especially) kinship terminologies were not simply representational--that kin terms had a core of culturally constructed, highly affective, and directive elements as well as a representational aspect (Schneider, 1965)--that the relevance of human emotionality and intentionality to the analysis of meaning in general became apparent to me (D'Andrade, 1976).

MEANING SYSTEMS VERSUS SYMBOL SYSTEMS

Throughout this paper the term meaning system has been used where the more conventional term symbol system may have been expected. The shift in terminology is intentional, based on a particular view about where meaning is and how it is organized.

The problem of "where meaning is" has been discussed by Douglas Hofstadter in his remarkable book, Godel, Escher, Bach (1979:158):

> The issue we are broaching is whether meaning can be said to be inherent in a message, or whether meaning is always manufactured by the interaction of a mind or a mechanism with a message. . . . In the latter case, meaning could not be said to be located in any single place, nor could it be said that a message has any universal, or objective, meaning, since each observer could bring its own meaning to each message. But in the former case, meaning would have both location and universality.

Hofstadter presents two different models of how messages are related to meaning, using as examples jukebox buttons versus a music record. He begins by pointing out that we feel quite comfortable with the idea that a record contains the same information as a piece of music, because we know (or trust) that there is an isomorphism, or one-to-one correspondence, between the physical characteristics of the groove patterns in the record and the sounds we hear.

When we push buttons on a jukebox and music comes out, we do not think that the buttons themselves contain the same information as the music we hear, even though something about the buttons produced the music. Since the characteristics of the buttons do not have a one-to-one correspondence to the characteristics of the music we hear, we realize that the musical information is not "in" the buttons but rather is triggered by the buttons. We reasonably assume that inside the machine is something that already contains the information necessary to produce the music (a very small musician, perhaps).

With respect to culture, this question is usually put in the following form: Where does one look for meaning--in culturally produced messages of various sorts or in the minds of the people who interpret these messages? If variations in the physical forms that make up the manifest message have the appropriate one-to-one correspondence to the meaning carried by the message, then the cultural analyst need only have the proper intelligence to detemine its meaning. But if the message is highly compacted and lacks the necessary isomorphism between variations in the physical signal and the meanings produced, then the message cannot be deciphered by the cultural analyst without recourse to the latent system already present in the mind of the decoder.

Some cultural messages appear to contain a great deal of internal structure, while other messages are closer to being triggers than to being records. The letters *d*, *o*, *g* are much more like jukebox buttons than a record of doggishness. On the other hand, the script of a play can have such a rich internal structure that in some ways it seems to be almost a recording of experience.

This problem of the location of meaning has been discussed at length by Schank and Abelson (1977), who have attempted to create computer programs that can understand such standard cultural messages as newspaper stories. They define understanding as the ability to answer questions about what happened in the story and to generate a paraphrase. In order to create programs to acccomplish this feat, Schank and Abelson have found it necessary to build into their programs knowledge about cultural roles, settings, goals, and event sequences. Say, for example, one reads the following story:

> Roger went to the restaurant. He ordered *coq au vin*. The waiter was surly and the table was right next to the cash register. Roger left a very small tip.

To answer questions such as:

What did Roger eat?
To whom did Roger give his order?
Where did Roger sit?
Did Roger like the restaurant?
Who was the tip for?

it is necessary to use cultural information (since none of the answers to these questions is explicity contained in the story) concerning the facts that one normally gets to eat what one orders, one normally gives ones order to the waiter, one normally sits at a table, one normally indicates satisfaction with food and service by the size of the tip, one is normally not satisfied if the waiter and the surroundings are not pleasant, and normally tables right next to the cash register are not pleasant. Furthermore, readers expect that, if something is not explicitly mentioned in a story, then what happened was what one would normally expect to happen. Without this additional information and the normality assumption, one could not answer the questions above and the story could not be understood.

In order to construct a program capable of understanding stories involving restaurants, Schank and Ableson built into their programs a complex restaurant script, containing information about physical props (tables, the menu, checks, money), roles (customer, waiter, cook, cashier, owner), entry conditions (the customer must have money, is usually hungry, and is attempting to obtain food), event sequences (entering, ordering, eating, exiting), and the causal relations between these items. In general it appears that understanding even simple messages in natural language requires considerable interaction between the information contained in the message and the information contained in the message processor. This conclusion seems reasonable for most things cultural--the meaning of rituals, games, myths, plays, texts, and other cultural forms is a complex product of what is contained in the representation and what the individual brings to the representation.

As a result of the interaction between what is contained in cultural messages and what is contained in the interpretative system of the mind, as a general rule one cannot locate cultural meanings in the message. Thus a distinction must be made between message and meaning. It is of some interest that the term symbol is ambiguous on exactly this point. That is, the term symbol can refer to either the physical thing that carries the meaning or to the meaning carried by the physical thing. Even when the term symbol is used with reference to something within the mind, it is typically refers to an internal image of some external form.

The ambiguity in the term symbol about whether the thing being referred to is something internal or external is related to the assumption that internal meanings are simply mental representations of the physical signals.

Thus if meanings are simply internal representations of external forms, then ambiguity about whether the external or internal forms are being referenced make no great difference.

The assumption that mental processing is the internal manipulation of representations of external signs may be correct. However, current work in cognitive psychology does not treat meanings as the representation of external forms, but as distinct entities having different principles of organization. Internal forms are typically called "schemata" and are considered to be composed of abstract proposition-like networks (Rumelhart, 1978). Since at this point we do not know which position is more accurate, it seems better to use distinct terms for internal meanings and external signals, thereby avoiding ambiguity and smuggled assumptions. The term meaning system has been used here throughout for mental structures and processes, rather than the term symbol.

One of the obvious but nevertheless remarkable facts about meaning systems is that interpretations of past messages can change the interpretative system itself, so that new messages are understood differently than they would have been had not the previous message occurred. This makes for a very flexible system. Added to the modifiability of meaning systems is the fact that people can produce messages and meanings that then react on the producer. The result of both these potentialities--modifiability and reflexiveness--is that people can change their own meaning systems--think things through and get things straight (or get themselves into a terrible muddle). However, there seem to be limits on how much self-induced change is possible, perhaps because at some point, for reasons yet unclear, without outside stimulation no new messages get produced.

Another notable property of meaning systems is that one can construct messages about messages and meanings about meanings. On the cultural level this phenomena is very extensive. For example, symbolic entities created by constitutive rules frequently become the topics of other constitutive rules, creating entities made of other culturally created entities. Thus theft requires the notion of property, sacrilege requires the notion of sacredness, etc.

Related to the process in which one meaning is the topic of another meaning is the framing of symbols (Bateson, 1972). In the public presentation of messages on printed signs, in books, at theatres, through ritual, on television, in group meetings, etc., messages are framed by context and by other messages telling what the original message is about (Goffman, 1974; MacAloon, 1981). To the extent that the recipients of these framed messages share the same relevant meaning systems, the meanings of these messages may be shared. Even when the meanings are not entirely shared, the fact that a number of people have access to the same physical messages creates the possibility of discussion and interpersonal negotiation concerning what was really meant (Cicourel, 1973; Holland, 1981.

In most human groups the communication of messages, both framed and unframed, is so frequent and redundant that it suggests the hypothesis that meaning systems need messages to keep themselves alive. Without relatively constant activation perhaps meaning systems disintegrate. While messaging, public and private, is an almost constant activity, the collecting of messages is not, I believe, the best way to start the study of a culture. The most fruitful place to start such study is with individual meaning systems. I hope that after this lengthy discussion of constitutive rules and culturally created objects I will not be taken as saying that culture is just a special sort of mentation. What I am trying to say is that the external signs, the public events, are too elliptical to serve as a good place to begin the search for organization and structure.

This is not to deny it is helpful to have a great deal of observation of what people in a culture do and say. Field observation is often necessary in order to understand what an informant is trying to describe and always necessary in order to understand that which informants cannot describe. However, field observation has become such an unquestioned virtue in anthropology that some calling to account might be valuable. It is relevant that when Metzger and Williams (1963) presented a series of demonstrations that they could obtain excellent ethnographic descriptions from informants without field observation, they were attacked with some vigor. The basis of the attack as I heard it was not that Metzger and Williams had gotten their descriptions wrong, but rather that such a procedure was not the way to do good ethnography. This objection presumes the very charge it seeks to prove. Objections to "white room ethnography," in which an informant is questioned in a situation removed from cultural context, may be based on a misperception of what is required for effective communication between informant and investigator.

THE STUDY OF CULTURE AS AN EXPERIMENTAL SCIENCE

Anthropology, it is said, is an observational science. Ethnographers in particular regard themselves as observers and consider participant observation to be the principal method of field research. Certainly, with regard to things like the network of social interaction or the operation of the economic system, there is little the investigator can do but try to observe and hope that what needs to be observed can be seen. The major alternative is to find someone in the society--an informant--who has observed the event that the ethnographer cannot observe and obtain the needed information through the informant's reports.

In the study of social interaction or economic systems there is little chance for an ethnographer to use experimental techniques, since ethnographers do not usually have the power to affect such systems. There have been attempts to create partial sociocultural systems in a laboratory environment and, by varying certain conditions, to study how changes in one variable affect other variables (Rose and Fenton, 1955). An interesting review of this type of work from an anthropological perspective has been presented by McFeat (1974), who also constructed in an ingenious manner a number of microsociocultural systems, then observed how differences in group size and kind of task influenced the development of cultural norms. But, however interesting, these miniature worlds are highly dependent and incomplete systems, better at demonstrating that we know how to produce a particular effect than for testing a hypothesis to see if it is really true.

It is not the case, however, that anthropologists cannot affect the people they study. Just to ask someone a question is to affect that person. What kind of event is this? Does it affect something cultural? Is asking a question an experiment? What is an experiment, and do anthropologists need them?

There is a rich literature on experimentation as a scientific technique (e.g., Carlsmith et al., 1976). The model presented in this literature contrasts sharply with the typical folk model of an experiment, with its white-coated scientists subjecting the object of their investigation to various kinds of unusual treatment with outlandish apparatus. Actually, an experiment does not require laboratories, or apparatus, or unusual conditions. An experiment requires three elements: first, an idea or propositions about certain things or events; second, a way to relate these things or events to operations and observations that can be made on something; and third, the power to determine which things at which times will have the operation done to them. Of course, some experiments are better than others: The original idea can be vague or uninteresting, the experimental operations carried out and the observations made may be only ambiguously related to the original idea, and the experimenter may have only limited power to select on what and when the operations will be done and only limited power to observe the effects. Despite such problems, which are common to all the sciences, what has been done is an experiment if the three conditions can be said to be present.

There are two major reasons for doing experiments. The first is that it may be difficult to find an opportunity to observe what one wants to observe. For example, suppose an investigator has a hypothesis that a particular set of plants will all have the same name. It is unlikely that the investigator will have the opportunity to observe several people naming each of these plants, since naming plants is something people do infrequently. An experiment, in which the investigator presents various people with the plants and induces them to name the plants, is a reasonable way to discover what cannot be observed naturally.

A second reason for doing an experiment is that a hypothesized relationship between X and Y is confounded by the fact that, in most natural settings, when X occurs, A, B, and C also occur, so that it is difficult to know if it is really X and Y that are related, or A and Y, or B and Y, etc. For example, an investigator might wonder how someone felt about his or her boss. If the person is always very respectful around the boss, it is difficult to know if this is because the person really respects the boss or because the person is afraid of the boss. By experiment through the presentation at various times of various questions and statements, an investigator can attempt to cut through the confounding conditions of the boss's presence with the employee's expression of feeling about the boss.

A question, put to discover if somebody believes something, or feels some way about something, or intends to do something, makes a very simple experimental operation. To the extent that one tests other people for the current representative, affective, and directive state of their meaning system, one has carried out an experiment. People constantly experiment on each other, using a variety of cultural forms: for example, direct and indirect questions, apparent disagreement intended to evoke deeply held commitment, apparent untruths intended to test whether a person really knows something, withholding acknowledging or back-channel responses to see if someone is saying something just for effect, etc. Lovers test lovers, believers test believers, knowers test knowers, people of purpose test the purposes of others, and ethnographers test informants. An important fact about meaning systems--idiosyncratic and cultural--is that they are accessible to this kind of experimentation.

Of course, many questions are asked not to make any specific test, but to simply try to find out something. Someone may ask, "How do you get to the post office?" just to find out how to get to the post office. Strictly speaking, no experiment has been done, because no specific hypothesis has been tested. However, even in the case of a simple question there are some implicit hypotheses about a person's meaning system being tested: that the question is understandable to the person, that the person knows where the post office is, that the person can respond with an understandable answer, that the person asked is likely to respond and to be truthful, etc. Thus a simple question does not have the goals of an experiment but does involve experimental tests of various sorts. In most cases ethnographers use a complex mixed strategy of experiment and simple question, starting with few assumptions about what the informant knows or feels and eventually building a theoretical structure about the informant's meaning systems.

There are a variety of experimental operations to test hypotheses about meaning systems other than the use of questions. The experimenter can present an object, create an event, or present the representation of some object or event, then observe the informant's reaction. The reaction observed can be something the person says, something the person chooses to do, how the person reacts emotionally, or how quickly the person responds, etc. The common procedure, however, is the question and answer format of natural language.

The preponderance of the question and answer format in cultural experiments is partially due to the ease with which questions can be asked and answers can be recorded. Another, even more compelling reason is that many of the important things in a culture, such as success or the soul, are entities that have no palpable form. To find out what an informant thinks or feels about symbolic things requires communication through a medium like natural language in which such things can be represented. In such cases discourse through natural language is almost the only means of investigation.

Given the type of question and answer experiment that anthropologists typically do, it is not surprising that rapport has emerged as one of the major research concerns. Unlike social psychologists, for instance, whose major difficulties in creating experiments involve the construction of conditions that have an appropriate correspondence to ordinary life, the primary experimental difficulty encountered by anthropologists is the development of the appropriate interpersonal conditions for verbal expression. Ordinary life gives people many reasons to hide how they think and feel and many special conventions about when and how and to whom various kinds of things may be said, so that the establishment of special rapport between investigator and informant becomes a major precondition for obtaining the type of communication anthropologists need. The task is to establish a relationship in which the informant understands the kinds of things the investigator wants to find out about and to trust the investigator enough to express things that might be punished in other contexts. And, since communication even under the best of circumstances tends to result in misunderstandings on both sides, the question and answer testing of informants' meaning systems is best done through a number of little experiments carried out over a long period of time. This kind of experimentation is typically informal, but it can also be undertaken in highly structured formats and combined with special techniques for the analysis of responses, such as multidimensional scaling (e.g., Gerber, 1976; Kirk and Burton, 1977; Roberts et al., 1981; Romney and D'Andrade, 1964; White, 1980).

From the perspective of the evolution of science, one would expect that each field would develop the experimental techniques that best fit the particular phenomena being studied. Something like this has happened in anthropology, in which there has developed in an unusually unselfconscious way a very special sort of experimentation, unrecognized as such, characterized by an emphasis on verbal interaction, subject-experimenter rapport, and successive testing. Related experimental methods have developed in linguisitics, clinical psychology and psychiatry, and in sociology--in all cases without the explicit recognition that what is being done involves the development of an experimental methodology.

The possibility of using experiments of certain sorts to investigate individual and cultural systems of meaning is an important part of the development of any sort of science based on meaning, because it makes feasible the investigation of individual interpretative systems. Through informant-based experimentation it is possible to investigate what would otherwise rarely be observed and to separate otherwise confounded conditions. This kind of testing makes it feasible to try to know enough about any person's systems of meaning to understand, and even predict, why a particular message is taken to mean one thing rather than another.

To return to a previous topic: As someone who observed Metzger and Williams's field procedures, I found that one of the most salient characteristics of their method was the development of a special kind of relationship with their informants, a relationship that was long-term, task-oriented, and marked by mutual respect. Informants understood the goals of the projects that they took part in and came to understand what it was that the anthropologists did and did not know. There was a professional quality about the interaction on both sides that was remarkable. In my view, it is not the color of the walls of the room that is important in working with informants or even the particulars of interviewing technique. What is important is the character of the investigator-informant relationship. Context is usually a social relationship--that is, the meanings people have for each other.

THE RELATIONSHIP BETWEEN MEANING SYSTEMS AND SOCIAL STRUCTURE

Issues about the nature of culture are intertwined with questions about the degree to which culture is shared and how culture is distinguished from social structure. In the 1940s and early 1950s, when culture was thought of primarily as the shared behavior distinctive of a social group, the problem arose that often things that seemed clearly cultural were not completely, or even generally, shared. For example, in American culture linear programming is important in engineering and business, but it is not a generally shared item of knowledge. John Roberts has pointed out that one of the functions of social organization is to create a division of labor in "who knows what." Roberts has also pointed out that societies differ in the way in which cultural information is integrated in the social process of decision making (Roberts, 1964). Marc Swartz has discussed the problems of the distribution of cultural understandings across social roles in detail and has proposed that cultures contain "linking understandings," in which those who occupy certain statuses share certain understandings about what other classes of persons are likely to know (Swartz and Jordan, 1976).

One of the basic things that meaning systems do for individuals is to guide their reactions and behavior. Given a systematic distribution of meaning systems across individuals--a system of systems--the reactions and behavior of groups of people can be guided in an organized and coordinated fashion. The concept of social structure appears to refer to the systems of systems of meanings. That is, social structure is usually defined as the distribution of rights and duties across status positions in a society. Each configuration of rights and duties is a culturally created entity, based on constitutive rules learned and passed on to succeeding generations. In this sense, social structure is one aspect of the organization of culture--the achievement of systematicity across persons through meanings.

THE RELATIONSHIP BETWEEN MEANING SYSTEMS AND SYSTEMS OF MATERIAL FLOW

In suggesting that culture and social structure are really composed of the same material, it may seem that what is being assumed is that all important human phenomena are basically meaning systems. This is not the case. There is a major class of human phenomena that is not organized as meaning systems, which I term material flow. By material flow I mean the movement of goods, services, messages, people, genes, diseases, and other potentially countable entities in space and time. These materials can be grouped into various classes, such as economic transaction, demographic change, interpersonal exchange of messages, ecosystem energy exchange, etc., and studied scientifically as systems with certain lawful properties. In the social sciences economics is the most developed of such disciplines, and its models have been widely extended to other kinds of systems of exchange.

Perhaps it is not surprising that those anthropologists whose major interest has been in the cross-cultural study of material flow were the most vehement in their rejection of the cognitively oriented view of culture proposed in the 1950s. This group, made up primarily of cultural evolutionists and cultural ecologists, treats culture as a system of socially transmitted standing behavior patterns through which communities adapt to their ecological settings--that is, as a kind of material flow of behaviors on a par with other kinds of material flow. This approach has the virtue of staying with what is most observable and maintaining a strong connection with the methods of the natural sciences. From the point of view of this paper, the problem with such an approach, as Roger Keesing has suggested, is that standing behavior patterns are influenced by so many variables (e.g., social crowding, climate, warfare, local geographic features, physical distribution of foodstuffs, prevalence of diseases, technological sophistication, plus meaning system characteristics) that there is little that can be said about such a phenomenon except that it seems so unstable that it is not likely to be worth studying (which in fact is what cultural materialists say about what they term culture). Keesing, following a long tradition, suggests that we use the term sociocultural system for the total system that includes behaviors, other types of material flow, and meanings, while reserving the term culture for meaning systems (Keesing, 1974:75).

An important issue is the way in which cultural meaning systems relate to systems of material flow--that is, to systems in which material and cultural objects move across time and space. Let us consider this relationship for two systems: the social structure, defined as systematic distribution of meaning systems across statuses, and the social exchange network, defined as a potentially observable flow of commands, services, goods, sentiments, etc., across persons. The social exchange network consists of rates by which various objects--which may be culturally constituted objects, like wealth or commands, or purely physical quantities, like bushels of wheat or pounds of iron--move from individual to individual. As defined here, the flow of such objects is not the same as the social structure, since groups with similar social structures can have very different social exchange networks, and similarity in the social exhange network does not necessarily mean that two groups have similar social structures. Thus, for example, groups with very similar rules of marriage (social structure) may vary widely when one counts actual marriages (social exchange network) because many conditions--such as the size of the various marrying groups, the age composition of groups, the distribution of wealth across groups, etc.--affect marriage rates.

The meanings that make up the social structure affect the flow of the social exchange network in numerous ways. For example, the various constructions concerning rights and duties expressed in norms of inheritance, rules of land tenure, wage scales, the norms for the ascription and achievement of statuses, the conventions of etiquette, etc., will all affect the flow of interaction, resources, commands, etc. Constitutive rules create a variety of types of people, occasions, and objects, which are linked to norms concerning the rights of certain types of people over certain kinds of objects on certain occasions. These norms are major determinants of a person's actions and reactions. Thus the meaning system directly affects the flow of things on which social life depends.

The causal relation is not one-way, however. As external conditions change, rates of various kinds of exchanges change, creating social opportunities and problems, which people adapt to with new norms and eventually new constitutive entities (Bailey, 1960). For the last 3,000 years or so human culture and society have been undergoing extremely rapid change. Sometimes it seems to me as if trying to study human culture in the 20th century is like trying to study the physics of moving bodies while living in the middle of an avalanche. Equilibrium conditions are rare, and what looks like stability is just the fact that most things are moving rapidly in roughly the same direction.

The fact that there are multiple two-way causal relationships between meaning systems and conditions of material flow has some strong consequences with regard to the possible kinds of analysis one can do with even the most carefully collected data. When experimentation is impossible, determination of the size of causal efects--how much change in one variable will affect change in another variable--is sometimes possible through correlational analysis. However, when there are feedback loops among the variables (e.g., A influences B, B influences C, C influences A), it is mathematically impossible, no matter how many data are collected, to estimate with any accuracy the degree of influence--unless one can assume that an equilibrium condition has been reached, so that the system is stable. In the language of path analysis, models with feedback loops are called "nonhierarchical models"--since causation does not always go in just one direction. David Kenny, in a text on methods of inferring causality from correlational data, states: "My own suspicion is that the strong assumption of equilibrium is sufficiently implausible to make nonhierarchical models generally impractical for the applied social scientist" (1979:105). In general, then, if one believes that technology influences ideology and that ideology influences technology, there are no data that can be analyzed mathematically to tell us whether technology is a more powerful causal variable than ideology, or the opposite--at least not until equilibrium conditions are found.

Thus causal priority debates, like the Whiting-Young debate concerning the causal priority of early experience versus factors of social organization on the severity of initiation ceremonies, are probably undecidable in principle. Rather than trying to find out which causal variables are the important ones, I believe that a more effective strategy for the human sciences, when experimentation is not possible, is to try to isolate patterns or configurations of variables that occur together frequently and have some stability over time, and to try to find which sets of patterns can change into other sets of patterns. Other social scientists have come to similar conclusions for less statistically motivated reasons.

While there may be great difficulty in determining the size of effects, there is little doubt that changes in systems of material flow do influence cultural meaning systems. However, the processes by which this influence takes place are not well agreed on. Since meaning systems are part of the human psyche, cultural meaning systems can be changed only through psychological processes. For example, changes in residence patterns appear to affect kinship terminologies--but for this to happen there must be some psychological process by which the change in people's experience leads to change in the conceptual classification of kin and the encoding of the new classification system into the language. Using one particular psychological theory of how people learn to make discriminations--a one-element stimulus-sampling theory--I developed a model that did a reasonable job of "predicting" kinship terminologies from features of social organization. Unfortunately, several years later, the particular psychological theory I had used was found to be too simplistic and is now generally considered inadequate to account for complex discrimination learning (D'Andrade, 1971). Over the past years the psychological theories of the process by which experience affects concept formation have changed continually and at this point do not seem close to resolution (Cole et al., 1981). It is sometimes discouraging to work in psychological anthropology in areas in which psychological theory is not well formulated. The conclusion I have come to is that psychological theory is most useful as a heuristic for exploring facts about the organization of culture and least useful as explanatory postulates.

THE RELATIONSHIP BETWEEN MEANING SYSTEMS AND PERSONALITY

Personality is another kind of system that is distinct from but related to cultural meaning systems. One formulation of the relationship between these two sets of systems is that ideas, values, and attitudes that are shared by a group are culture, but these same things, if idiosyncratic, are personality. A different formulation is that those ideas that an individual has to know to behave appropriately as a member of society are culture, while values and attitudes are personality.

There are problems with both these formulations. With regard to the "shared learning is culture, idiosyncratic learning is personality" formulation, most learned things are somewhat shared, but nothing is ever shared completely, so that everything people learn ends up being a little bit personality and a little bit culture. With regard to the "ideas that one has to know to behave appropriately are culture, values and attitudes are personality" formulation, most personality theorists would want to include some of things one has to know to behave appropriately as part of individual personality, and some cultural theorists would want to include some values and attitudes as part of culture.

The basic drawback of these content-based formulations can be illustrated by imagining that chemists and biologists had decided to divide the world into physical objects that are chemical and physical objects that are biological. One could imagine interesting arguments about whether proteins are biological or membranes are chemical.

Rather than a content-based formulation, it seems more useful to consider items of human learning as either culture or personality, depending on how they are placed within a system of relationships and processes (Kracke, 1981). In the study of culture these relationships and processes involve the adaptation of the groups of people to their environment and to each other through systems of meaning. In the study of personality these relationships and processes involve the organization of behavior, impulse, affect, and thought around the drives of the individual. If one considers personality and culture to be open systems that are linked together, then there must be items belonging to both systems that form the links. For example, ideas concerning success may, for most Americans, be not only part of their cultural meaning systems, but also a part of their motivational system, and thus play a part in both culture and personality.

The problems concerning the relation between culture and personality raise the issue of the relation between culture and experience. Humans experience a complex universe, composed of perceptions, memories, thoughts, and fantasies about social and physical events. Only a part of any one person's experience is shaped by or represented through particular systems of cultural meaning. Modern American culture has much more to say about the experience of being young than the experience of being old, for example. There is always interplay between the world of experience and cultural meanings; in some cases cultural meanings have the potential of giving form and depth to private experience, in some cases cultural meanings may conflict with and distort the individual's experiences, and in other cases there may be no relation established by the individual between particular experiences and cultural meanings. Just as there is a dynamic between cultural meanings and systems of materials flow that creates a potential for change, so, too, there is a dynamic between cultural meanings and private experience that also creates a potential for change.

CULTURE DEFINED

The oldest terminological wrangle in anthropology is over the term culture. Some of the problem seems to come from the fact that the term has a sense both as a process (the "passing on" of what has been learned before to succeeding generations) and as a particular class of things ("shared knowledge," for example). One might think that these two aspects of the terms could coexist quite neatly if the process were used to define content. In such a definition, culture would be whatever it is that is passed on through learning to succeeding generations. The difficulty with this solution is that many things are passed on, not all of which most anthropologists would want to consider culture. For example, oedipal complexes are learned and shared widely (even in the Trobriand Islands; see Spiro, 1981) but would not usually be considered to be culture by most anthropologists, since such complexes are an indirect, unintended, and unrecognized consequence of the learning of other things. A second strategy is to define culture as having a particular content. The problem with this solution is that there are different kinds of content--which should get to be called culture?

Such terminological quarreling might seem to be academic foolishness. However, most of the battles about the nature of culture have been generally enlightening, perhaps because they make explicit our assumptions about what is out there, and perhaps because as our assumptions about what is out there become explicit, we find that there are more kinds of things out there than we had thought.

Technically, anthropological "definitions" of culture are not definitions at all: According to Suppes, for example, a definition should not introduce a new axiom or premise in a theory or strengthen the theory in any substantive way (1957:153). Technically, a definition should be a paraphrase that maintains the truth or falsity of statements in the theory when substituted for the word defined. But to produce substitutable paraphrases for already existing propositions has not been the goal of those who have attempted to define culture. Rather, their attempts have been to describe what is out there--that is, to formulate substantive propositions about one aspect of the human world.

There are, at present, at least three major views about the nature of culture. One is a notion of culture as knowledge, as the accumulation of information. According to this view, culture can and does accumulate and does not need to be shared if the distribution of knowledge is such that the proper "linking understandings" are maintained. The amount of information in the total cultural pool of knowledge is very large--even for simple societies my estimate is that there are between several hundred thousand to several million "chunks" of information in the total pool (D'Andrade, 1981). Furthermore, in this view culture is not highly integrated; the knowledge concerning what to do about illness has no particular connection or relation to the knowledge needed to build houses, for example.

A second view is that culture consists of "conceptual structures" that create the central reality of a people, so that they "inhabit the world they imagine" (Geertz, 1981), or, according to Schneider, "elements which are defined and differentiated in a particular society as representing reality--not simply social reality, but the total reality of life within which human beings live and die" (1976:206). According to this view, culture is not just shared, it is intersubjectively shared, so that everyone assumes that others see the same things they see. In this view culture does not particularly accumulate, any more than the grammar of a language accumulates, and the total size of a culture with respect to information chunks is relatively small, if one can speak of size at all. The entire system appears to be tightly interrelated but not necessarily without contradictions.

A third view of the nature of culture falls between the "culture as knowledge" and the "culture as constructed reality" positions. It treats culture and society as almost the same thing--something made up of institutions, such as the family, the market, the farm, the church, the village, etc.--that is, systems or clusters of norms defining the roles attached to various sets of statuses. For Nadel, for example, these clusters of norms, if analyzed from the "who does what to whom" perspective, constitute social structure, and if analyzed from the "how one activity relates to another activity" perspective, constitute culture (Nadel, 1951). For Schneider, on the other hand, these clusters of norms also fall between the "knowledge" position and the "constructed reality" position with respect to accumulation, size, and integration. Accumulation occurs, but relatively slowly; the size of the body of information that must be learned is very large, but not thousands of thousands of chunks; and the degree of integration is important, but problematic.

Given the position taken in this paper, all three are views of cultural meaning systems. The difference between the views is in the prominence given to the various functions of meaning--to the directive function for the "norms and institutions" view, to the representative function for the "knowledge" view, and to the potential of systems of meaning to create entities for the "constructed reality" view. While there is a certain amount of differentiation among symbols and meanings--some seem primarily representational (e.g., propositions about farming), some seem primarily directive (e.g., propositions about how to raise children), and some seem primarily reality-constructing (e.g., propositions about what counts as what)--this differentiation is not sharp, and much of the apparent difference is in the conceptual framework of the analyst, not in what things mean to the individuals involved.

In summary, the position taken in this paper treats culture as consisting of learned systems of meaning, communicated by means of natural language and other symbol systems, having representational, directive, and affective functions, and capable of creating purely cultural entities and particular senses of reality. Through these systems of meaning groups of people adapt to their environment and structure interpersonal activities. Cultural meaning systems affect and are affected by the various systems of material flow, such as the flow of goods and services, and the interpersonal network of commands and requests. Cultural meaning systems are linked to personality systems through the sharing of specific items that function in both systems for particular individuals. Various aspects of cultural meaning systems are differentially distributed across persons and statuses, creating institutions such as family, market, nation, etc., which constitute social structure. Analytically, cultural meaning systems can be treated as a very large diversified pool of knowledge, or as partially shared clusters of norms, or as intersubjectively shared, symbolically created realities. On the individual level, however, the actual meanings and messages that people learn, encounter, and produce are typically not divided into separate classes of items that can be labeled knowledge, norm, or reality, but rather form multifunctional complexes of constructs, organized in interlocking hierarchical structures, which are simultaneously constructive, representive, evocative, and directive.

REFERENCES

Anscombe, G. E. M.
1958 "On brute facts." Analysis 18:3.

Austin, J. L.
1962 How to Do Things With Words. Oxford, England: Oxford University Press.

Bailey, F. G.
1960 Tribe, Caste, and Nation. Manchester, England: Manchester University Press.

Bateson, Gregory
1972 "A theory of play and fantasy." In Steps to an Ecology of Mind. New York: Ballantine.

Carlsmith, J. M., P. C. Ellsworth, and E. Aronson
1976 Methods of Research in Social Psychology. Reading, Mass.: Addison-Wesley.

Casson, Ronald W.
1981 Language, Culture, and Cognition. New York: Macmillan.
Chomsky, Noam
1957 Syntactic Structures. The Hague: Mouton.
Cicourel, Aaron
1968 The Social Organization of Juvenile Justice. New York: Wiley.
1973 Cognitive Sociology: Language and Meaning in Social Interaction. London: Penguin.
Cole, Michael, and the Laboratory of Comparative Human Cognition
1981 "Intelligence as culture practice." In W. Kessen, ed., Carmichael's Handbook of Child Psychology. New York: Wiley.
D'Andrade, Roy
1971 "Procedures for predicting kinship terminologies from features of social organization." In Paul Kay, ed., Explorations in Mathematical Anthropology, Cambridge, Mass.: MIT Press.
1976 "A propositional analysis of U.S. American beliefs about illness." In Keith Basso and Henry Selby, eds., Meaning in Anthropology. Albuquerque: University of New Mexico Press.
1981 "The cultural part of cognition." Cognitive Science 5:179-195.
D'Andrade, Roy, and Michael Egan
1974 "The colors of emotion." American Ethnologist 1:49-63.
Friedrich, Paul
1977 "Sanity and the myth of honor: the problem of Achilles." Ethos 5(3):281-305.
Geertz, Clifford
1973 "The growth of culture and the evolution of mind." In The Interpretation of Culture. New York: Basic Books. (Originally published in J. Scher, ed., Theories of the Mind, 1962.)
1981 "The way we think how: towards an ethnography of modern thought." Bicentennial Address of the American Academy of Arts and Sciences.
Geoghegan, William
1971 "Information processing systems in culture." In Paul Kay, ed. Explorations in Mathematical Anthropology. Cambridge, Mass.: MIT Press.
Gerber, Eleanor
1975 The Cultural Patterning of Emotion in Samoa. Ph.D. dissertation. University of California, San Diego.

Goffman, Erving
1974 Frame Analysis. New York: Harper.

Goodenough, Ward H.
1957 "Cultural anthropology and linguistics." In Paul Garvin, ed., Report of the Seventh Annual Round Table Meeting on Linguistics and Language Study. Georgetown University Monograph Series, Language and Linguistics 9. Washington, D.C.: Georgetown University.

Hofstadter, Douglas R.
1979 Godel, Escher, Bach: An Eternal Golden Braid. New York: Random House.

Holland, D.
1981 "Samoan folk knowledge of mental disorders." In A. Marsella and Geoff White, eds., Cultural Conception of Mental Health and Therapy. Unpublished manuscript.

Keesing, Roger M.
1974 "Theories of culture." In R. Casson, ed., Language, Culture and Cognition. New York: Macmillan. (Reprinted in 1981.)

Kenny, David
1979 Correlations and Causality. New York: Wiley.

Kirk, Lorraine, and Michael Burton
1976 "Physical versus semantic classification of non-verbal forms: a cross cultural experiment." Semiotica 17(4):315-331.

Kracke, Waud H.
1981 "The complementarity of social and psychological regularities: leadership as a mediating phenomenon." Ethos 8(4):273-285.

Lakoff, George, and Mark Johnson
1980 Metaphors We Live By. Chicago: University of Chicago Press.

Levy, Robert I.
1981 Embedded and Marked Symbolism. Unpublished paper.

MacAloon, John J.
1981 Olympic Games and the Theory of Spectacle in Modern Society. Unpublished paper.

McFeat, Tom
1974 Small Group Cultures. Toronto: Pergamon Press.

Metzger, Duane, and G. Williams
1963 "A formal ethnographic analysis of Tenejapa Ladino weddings." American Anthropologist 64.

Much, Nancy C., and Richard A. Shweder
1978 "Speaking of rules: the analysis of culture in

breach." In William Damon, ed., New Directions for Child Development. San Francisco: Jossey-Bass.

Nadel, S. F.
1951 The Foundation of Social Anthropology. London.

Osgood, Charles E., W. H. May, and M. S. Miron
1975 Cross Cultural Universals of Affective Meaning. Urbana, Ill.: University of Illinois Press.

Quinn, Naomi
1975 "Decision model of social structure." American Ethologist 2:19-45.
1980 "Commitment" in American Marriage: Analysis of a Key Word. Paper presented at the meeting of the American Anthropological Association, Washington, D.C.

Rawls, John
1955 "Two concepts of rules." Philosophical Review 64:3-22.

Reddy, Michael
1976 "The conduit metaphor." In A. Ortony, ed., Metaphor and Thought. Cambridge, England: Cambridge University Press.

Roberts, John M.
1964 "The self management of cultures." In W. Goodenough, ed., Explorations in Cultural Anthropology: Essays in Honor of George Peter Murdock. New York: McGraw-Hill.

Roberts, John M., Garry E. Chick, Marian Stephenson, and Laurel Lee Hyde
1981 "Inferred categories for tennis play: a limited semantic analysis." In Alyee B. Cheska, ed., Play as Context. West Point, N.Y.: Leisure Press.

Romney, A. K., and Roy d'Andrade
1964 "Cognitive aspects of English kin terms." American Anthropologist 66:146-170.

Rose, Edward, and William Fenton
1955 "Experimental histories of culture." American Sociological Review 20(4):383-392.

Rumelhart, David E.
1978 "Schemata: the building blocks of cognition." In R. Spiro, B. Bruce, and W. Brewer, eds., Theoretical Issues in Reading Comprehension. Hillsdale, N.J.: Lawrence Erlbaum.

Schneider, David
1965 "American kin terms and terms for kinsmen: a critique of Goodenough's componential analysis

of Yankee terminology." American Anthropologist 67:part 2.

1968 American Kinship: A Cultural Account. Englewood Cliffs, N.J.: Prentice-Hall.

1976 "Notes toward a theory of culture." In Keith Basso and Henry Selby, eds., Meaning in Anthropology. Albuquerque: University of New Mexico Press.

Schank, Roger, and Robert Abelson

1977 Scripts, Plans, Goals, and Understanding. Hillsdale, N.J.: Lawrence Erlbaum.

Searle, John R.

1969 Speech Acts: An Essay in the Philosophy of Language. Cambridge, England: Cambridge University Press.

1978 "A classification of ellocutionary acts." Language and Society 5:1-23.

Shweder, Richard A.

1981 Anthropology's Romantic Rebellion Against the Enlightenment; or, There's More to Thinking Than Reason and Evidence. Paper presented at meeeting of the American Anthropological Association, Toronto, January.

Spiro, Melford E.

1961 "Social systems, personality, and functional analysis." In Bert Kaplan, ed., Studying Personality Cross-Culturally. Evanston, Ill.: Row, Peterson.

1981 Oedipus in the Trobriands: The Making of a Scientific Myth. Unpublished paper.

Suppes, Patrick

1957 Introduction to Logic. Princeton, N.J.: Van Nostrand.

Swartz, Marc J., and David K. Jordan

1976 Anthropology: Perspectives on Humanity. New York: Wiley.

Vendler, Zeno

1967 Linguistics in Philosophy. Ithaca, N.Y.: Cornell University Press.

1972 Res Cognitans: An Essay in Rational Psychology. Ithaca, N.Y.: Cornell University Press.

White, Geoffrey M.

1980 "Conceptual universals in interpersonal language." American Antrhoplogist 82(4):759-781.

The Life-Span Perspective in Social Science Research

David L. Featherman

INTRODUCTION

Over the last decade a convergence of orientations on human development has emerged within several social and behavioral sciences. Known as the life-span perspective on development and behavior, the essence of this approach is that developmental changes in human behavior, which occur from conception to death, and which arise from a matrix of biological, psychological, social, historical, and evolutionary influences and from their timing across people's lives. Scholarly and popular interest in the themes of this perspective have been intense since 1970. The popular press, for example, has absorbed the ideas of the mid-life crisis and life-cycle passages; many articles have been written about the implications of changes in life-style, such as dual careers, the empty-nest phase of marital relationships, greater exposure to chronic disease with increases in longevity, and the return of adults and senior citizens to college and university classrooms. In the academic community evidence of the specialization of study is the growth of life-span developmental psychology, the publication of monograph series on life-span processes, and the organization of interdisciplinary conferences by private and federal agencies that have explored both aging and development as lifelong processes.

This paper was commissioned by the Social Science Research Council for The National Science Foundation's Five-Year Outlook on Science and Technology:1981.

BASIC THEMES AND PROPOSITIONS

The multidisciplinary study of life-span development is not yet based on a coherent theory with explicit hypotheses that can be addressed in empirical research. Instead, the recognized consensus is more of a model, paradigm, or world view, but the paradigmatic themes that have emerged have already generated a few propositions. These themes and premises reflect a reinterpretation of old evidence about child and adolescent growth, new findings from longitudinal research among the lives of several birth cohorts now entering late adulthood, and intervention studies of the aged. The thematic statements are challenges to conventional thinking and guides to future research. They can be summarized succinctly as follows:

1. Developmental changes occur over the entire course of life; they are synonymous with aging in the broadest sense. Aging is not limited to any particular time of life; neither is development.
2. Developmental changes in the course of aging reflect biological, social, psychological, physical, and historical events.
3. The multiple determinants of constancy and change in behavior and personality express their influences interactively and cumulatively, defining life event or life history trajectories.
4. Individuals are agents in their own development. Life histories are transactional products of the dialectics among the multiple determinants of development and the motivated, selectively responding individual. Generalizations across individuals about constancies in human development, especially throughout the last half of life, are few and difficult to formulate.
5. Each new birth cohort ages through a potentially different trajectory of life events, the product of sociohistorical change and individual reactions to it. Historically constant generalizations about developmental changes in aging are fewer or greater as a function of the pace and direction of sociohistorical changes, either evolutionary or revolutionary.
6. Intervention efforts on behalf of the aged are effective in changing the course of development, even as they are in the young. Behavior and personality apparently remain more malleable throughout life than is apparent in common contemporary social and subcultural settings. The apparent plasticity of manifest patterns of

development--untapped reserve or potential--suggests a rethinking of research paradigms and social policies that are predicated on conventional, essentially static models of aging and of stable and universal stages of development from childhood through old age.

These themes and propositions are developed in recent multidisciplinary, edited collections by P. B. Baltes (1978), P. B. Baltes and O. J. Brim (1979, 1981), Riley (1979), and Brim and Kagan (1980). For a detailed discussion of the empirical bases of these themes, see Featherman (forthcoming).

HISTORICAL OVERVIEW

The themes and propositions that provide the focus for these developments have a long history within the social and behavioral sciences in the United States and abroad. Only in the last 10 years, however, have they begun to achieve a scientific base, in some instances through independent work in two or more academic disciplines. Through repeated consideration of the interrelatedness of change across individuals' life-cycles and social or institutional change, the scientific activity of the past decade has been extraordinarily productive in sharpening concepts, suggesting hypotheses and empirical research, and promoting long-range programmatic study of the relationship between the individual and social change within the boundaries of academic disciplines (e.g., psychology, sociology, history, and anthropology). It has also been a period of sustained discussions across disciplinary boundaries and of acknowledged convergences in certain points of view, optimal research methodologies, and intervention strategies. Some have even speculated that what is emerging from the current discussions and programs of longitudinal research may provide the basis for a new academic discipline, which would amalgamate and synthesize the traditions of its parent disciplines (Riley, 1981: 340).

To place these achievements in perspective and provide a means of assessing their implications, this paper begins with a brief historical review, organized by academic disciplines to highlight specific contributions. This organization does not emphasize the intense effort of the last decade to transcend disciplinary backgrounds and to explore commonalities in life-span concepts and methods.

The several disciplines do, however, enrich these explorations of the life-span approach with their unique perspectives.

Psychology

Psychologists have been among the most prominent proponents of a life-span approach to the study of human development in the United States. One reason is that the fundamental research agenda of psychologists is the study of an individual's behavior over time (i.e., P. B. Baltes et al., 1980b). Although other disciplines, such as sociology and history, also study change, the unit of analysis is frequently not the individual. This difference is fundamental to the life-span approach, as is apparent in two other reasons for the intellectual centrality of psychology to this area: One is the recognized importance of the biological or organic substrata of behavior (e.g., Rodin, 1980). Another is the theme of the active organism as an agent in its own development or as the instigator of change (e.g., Lerner and Busch-Rossnagel, forthcoming).

A life-span orientation to the study of human development has a long range in psychology. Reinert (1979) traces these roots to the philosophical writings and autobiographical reflections of Aristotle, Democritus, Augustine, and other classical philosophers. Perhaps the most important figure during what Reinert calls the "preliminary period" was Johann Nikolas Tetens (1736-1806), a German philosopher and experimenter in psychology. Tetens emphasized the importance of identifying general psychological laws of behavior through naturalistic observation. He placed the study of "developmental courses" of lives within the span from conception to death and stressed the importance of social and cultural conditions for development.

During the "formative period" (between the late 18th and late 19th centuries), two Europeans played key roles. Freidrich August Carus (1770-1808) saw the course of human development as a series of stages, each preparing the individual for what comes later in life. He strived to establish a "general age-oriented science," but one in which age was not assumed to be a causal variable per se and in which historical context was an important facet of age-related behaviors. Carus's conception of development was not supplanted as an intellectual guidepost for over a century (Reinert, 1979:220). The other key figure was Adolphe Quetelet (1796-1874), whose methodological contributions were as important as his substantive ones. He pioneered the use of cross-sectional analysis for the study of age-specific individual differences in development and emphasized the necessity of longitudinal designs for the analysis of intraindividual change (and interindividual differences in these changes). Quetelet identified the developmental importance of critical periods in the life-span and showed empirically that historical events or periods could be associated with change in developmental functions.

The first widely recognized American textbook that espoused a life-span view, Life: A Psychological Survey by Pressey, Janney, and Kuhlen, was published in 1939. It was not until the 1960s that the rubric life-span was commonly used to differentiate an "age-irrelevant" developmental psychology from the study of developmental processes in the age-specific academic specialties such as child development or adolescent psychology. Research and teaching programs at the University of Chicago (i.e., Havighurst, Neugarten) and the University of Bonn (i.e., Thomae) were forerunners. In the 1960s and 1970s, several symposia at the University of West Virginia advanced the conceptual and methodological frontiers of the life-span orientation, and the symposia monographs helped to institutionalize it (P. B. Baltes and K. W. Schaie, 1973; Datan and Ginsberg, 1975; Datan and Reese, 1977; Goulet and Baltes, 1970; Nesselroade and Reese, 1973).

It is likely that advancement of the life-span approach was retarded by several related factors. One was the intellectual dominance of child-focused developmental psychology, which placed the primary causes of lifelong behavioral tendencies within the childhood years. Associated with the hegemony of child psychology were conceptions of development--so-called biological growth models--that limited the scope of developmental thinking. These conceptions assumed an end state (i.e., maturity) toward which developmental changes and the universality of developmental patterns moved across individuals. Placed against many behaviors and psychological characteristics of children and adolescents undergoing apparently rapid quantitative and qualitative changes (e.g., bodily growth, sexual maturation), these modes of thinking about general development had considerable analytical utility. At the same time, the biological growth model did not fit adult behavioral changes as well as it did those of children.

Some psychologists argued that adults do not develop in the same sense as children and that the domain of developmental psychology (equating behavioral development with restrictive, biological analogues) should be bounded by the age-related stabilization of behavioral changes in early adulthood (e.g., Flavell, 1970).

A shift in the definition and causal explanation of development occured in the 1960s, and one of the major influences on this shift came from longitudinal studies of children initiated prior to World War II. Studies undertaken by the Institute of Human Development at Berkeley and the Fels Institute in Ohio were beginning to report the developmental trajectories for subjects who were reaching adulthood in the 1960s and 1970s. The availability of these extensive longitudinal data motivated psychologists to address developmental issues over longer segments of life than had been customary. And the empirical findings about the patterns of stability and change, predictability and discontinuity, fueled a controversy over fundamental perspectives and definitions (compare Block, 1971; Kagan, 1980; Sears, 1980; Thomae, 1979). The data did not conform to conventional metatheories that had been relatively unchallenged and productive when applied to developmental issues in childhood and adolescence, and strong differences in interpretation arose.

An even stronger impetus to the development of life-span orientations came from gerontologists. Gerontology as a field of inquiry was itself evolving during the 1950s and 1960s, when contributors such as Robert Havighurst, Bernice Neugarten, and James Birren began to articulate a psychology of aging. Gerontology, like other age-specific developmental specializations, was influenced by the dominance of the biological growth model, although its main focus was on senescence and decline rather than on growth and differentation. The orientation of the psychological gerontologists was by necessity much broader than that of their counterparts who studied children. The gerontologists placed biological aging and behavioral changes in older adults within the context of cumulative life histories, linking contemporaneous changes with sequences and events over the entire life-span. They became as concerned with the emergence of novel behaviors in their older subjects as they were with ones that could be construed as outgrowths of predisposing earlier experiences. They observed large differences in the courses of aging and in behaviors among adults and they sought explanation of this variation in historical circumstances and in chance occurrences, such as illnesses, accidents, and births of grandchildren, that are more individualized in their impact. In short, the psychology of aging encouraged a more contextual, historical approach to behavioral change and development, a tendency encouraged by the interdisciplinary character of gerontology as a field of inquiry. (An excellent example of such a contextual approach to development is Bronfenbrenner, 1979.)

As psychologists enlarged the concept of development to incorporate behavioral changes that were not adequately described by the biological growth model, they turned increasingly to the work of sociologists and historians, for which the context of behavior across the life-span was a central focus. The complementarity of the approaches encouraged cross-disciplinary collaboration and the emergence of common perspectives; one illustration is the series Life-Span Development and Behavior, begun in 1978 and edited currently by Paul B. Baltes, a developmental psychologist, and O. G. Brim, Jr., a sociologist. This work is but one of several recent products of the emerging multidisciplinary interest in life-span processes.

Sociology

Unlike psychology, no new life-span specialty has formed within sociology; however, this may simply show that sociologists have not felt the need to "rediscover" this orientation or to differentiate it from their long-standing intellectual concerns with socialization (e.g., Brim, 1966; Goslin, 1969) or age-differentiation (e.g., Cain, 1959, 1964; Elder, 1975; Riley et al., 1972). Indeed, sociological scholarship on aging from birth to death, amplified by contact with psychologists and historians over the last decade, has only begun to tap the intellectual traditions of a latent life-span approach within the subdisciplines of social psychology, social organization, and social demography.

Progress toward a comprehensive integration of these traditions into a sociology of the life course has moved along three loosely coordinated fronts: the sociology of age stratification, socialization, and the social demography of the family. Among the more programmatic contributions is the three-volume work of Matilda W. Riley and her associates, Aging and Society (1968, 1969, 1972). They reviewed the nearly inchoate literature on age-related behaviors of adults in the middle and later years, identified a series of conceptual and methodological flaws, and presented a conceptual model for integrating and promoting multidisciplinary research on aging. The age stratification model of Riley et al. views persons as aging biologically and psychologically over the entire span of life within an evolving social context.

The age stratification model emphasizes the social aspects of aging in three respects. First, lifelong aging reflects sequences of social positions, or trajectories of social roles and associated statuses and perquisites, that have age-related features. For example, schooling in childhood and adolescence typically precedes labor force entrance for American males, followed by marriage or parenthood. Second, social positions mold and reformulate behavior and personality as a person learns to perform and moves through sequences of positions. Third, both the patterns of biological aging and the sequences of age-related roles themselves are subject to change, for example, by secular improvements in nutrition or health care, by historical events such as wars and depressions, or by evolving institutional changes like industrialization, retirement legislation, or the diffusion of television. This last aspect implies that each birth cohort ages in a potentially unique way. It also implies that in any given year in which the age differences in a behavior or attitude are measured across a society's population two kinds of differences are manifest: those arising from "normal" aging--i.e., developmental patterns that are relatively invariant across historical time--and those reflecting the impact of unique historical experiences of persons born in different years. Both of these manifestations are abstractions, for what is actually observed is an interaction between cohort experience and age. Finally, it implies that one way in which society's institutions change is through the succession of cohorts--the replacement of older persons by younger ones as carriers of unique cumulative experience and outlook, with a unique developmental history. Thus, social change and individual (developmental) change are reciprocally and dynamically interrelated.

The conceptual framework developed by Riley and others reflects and summarizes a broad scope of earlier literature on aging and life course processes. For example, sociologists and social demographers such as Cain (1964), Ryder (1965), and (much earlier) Karl Mannheim (1952) elaborated the significance of generational and cohort succession and flows through society as major mechanisms of social change and ideological conflict. Sorokin, (1941, 1947), Parsons (1942), and Eisenstadt (1956) focused on the age structure of society and puzzled over the question of why and when age is used by a society as a means of sorting people into positions and as a device for allocating goods and service--much as social class or sex serve these societal purposes. Subsequently, others have elaborated the implications of ordered and disordered flows of birth cohorts for the management of potential conflict by institutions such as schools and the economy (e.g., Waring, 1975). Some (e.g., Foner and Kertzer, 1978) have attended to other cultural settings, principally the "age-set" societies of East Africa, in which individuals not only experience life events in an age-related sequence but also belong to a named group of individuals of like age who make "life-stage" transitions as a group. These cross-societal comparisons have suggested culturally specific as well as universal patterns of aging and human development. This line of inquiry has provided a natural link to an intensifying interest among anthropologists in aging as a lifelong process (see Keith, 1980).

Finally, the effort by Riley, Neugarten (i.e., Neugarten and Datan, 1973), and other sociologists to articulate the connection between the age-graded features of biological and sociocultural life events, on one hand, and historical and institutional change, on the other, has linked the life-span approach to other active areas of research on the society's organizing structures. For example, the age stratification model enables sociologists and anthropologists who study macrosocial or institutional patterns of socioeconomic inequality and mobility to integrate their research with that of social psychologists and developmentalists studying microsocial change--i.e., change over the course of individuals' lives (see the section on social inequality and stratification below). These types of integrations have fostered the multilevel character of life-span research.

A second tradition that is being tapped in integrating a sociology of the life course is socialization research. Socialization is a term used widely since the 1940s to refer to the complex processes whereby an individual learns and modifies the behaviors, values, and emotions that are deemed appropriate by the community and the larger society. In the mid-1960s, a conceptual monograph by O. G. Brim, Jr., placed socialization within an explicit life-span orientation (Brim, 1966). It differentiated the socialization patterns of children from those experienced by adults throughout the life-cycle in terms of the demands placed on the learner, what was to be learned, and the learner's role in the learning process. Brim thus reflected an orientation that had lain dormant in social psychological writings about culture and personality since the seminal research of W. I. Thomas (Thomas, 1909; Thomas and Znaniecki, 1918) and John Dollard (1935) and even tapped roots from the late 19th century (e.g., Giddings, 1897).

At the same time, John Clausen, a sociologist associated with the longitudinal Berkely Child Guidance Study, elaborated a concept of lifelong behavioral and personality change through a succession of shifts in roles and role sequences that constituted the course of adult life for persons in different social clases (Clausen, 1972). Clausen (1968) also edited a collection of essays on the life-span nature of socialization, which set the stage for several other empirical studies of adult socialization during the 1970s (e.g., Kohn, 1969; Kohn and Schooler, 1973, 1978). Further research on adult socialization highlighted the middle years (ages 40-60), an area neglected by the focus of much developmental research on children and the focus of gerontology on old age. Likewise, the focus on transitions between work and nonwork roles and attention to phenomena such as the mid-life crisis emphasized the continued development of the individual throughout life and gave conceptual coherence to the study of adult socialization.

By and large, however, socialization research was not cast within a life-span perspective with sustained intensity by sociologists. American sociologists have viewed personality development as part of the agenda of childhood and adolescent socialization--to reflect the template of social norms, cultural values, and sanctioned behaviors of the collectivity as these are impressed on preadults. The dominant tendency has been to think of individual behavior in terms of the functioning and persistence of society--socializing the child to assume his or her place in the social order--and therefore to emphasize commonalities in childrearing patterns that reflect themes of continuity (e.g., intergenerational persistence of values) and of consistency (e.g., correspondence between the needs or functions of society and the personalities or attributes of a society's population). (See Brim, 1980; Brim and Ryff, 1980; DiRenzo, 1977; and Elder, 1975 for insight into these trends within social psychology as a subdiscipline of sociology.) In part this interrupted and episodic intellectual development is a consequence of sociology's lack of a well-articulated concept of personality and its inattention to mechanisms of personality change (development). In addition, socialization research until the 1970s focused primarily on role transitions that were substantially correlated with chronological ages--e.g., the "disengagement" of the retired elderly, entry into parenthood--lending it a character not unlike the age-specialized approaches of developmental psychology up to World War II. The intensification of life-span approaches is undoubtedly tied to the contact with developmental psychologists in search of an understanding of context as well as to other institutionalizing influences.

The third area in which a life-span orientation is taking form in sociology lies at the interface of social demography and social history. It is a long-standing practice for social demographers, whether sociologists, historians, or economists, to study birth cohorts as a way of gaining insight into the impact of secular change on fertility behaviors, mortality and health, and migration--that is, on behaviors affecting the demographic equation. The last decade saw the diffusion of the cohort method of analysis and the application of a behavioral approach to the study of change into social or family history (e.g., Vinovskis, 1977). At the same time, sociologists studying the dynamic relation of social structure to personality had begun to probe its historical dimension. For example, Alex Inkeles and David Smith (1974) explored the complexities of contact between premodern man and the institutions of industrial society (e.g., factories and schools) to gain an understanding of how the personality traits we associate with modern mankind are elicited and selectively promoted. Glen Elder (1980) contrasted two cohorts of California children who had experienced the economic and social disruptions of the Great Depression at different points in their development, noting the importance of historical events as major influences on personality differences between groups born at different times but also among persons within a given cohort. These simultaneous intellectual developments came together in a series of cross-disciplinary research projects and monographs (Demos and Boocock, 1978; Hareven, 1978) in which scholars viewed personality and family structure as both a reflection and a modifier of historical change during the 19th century.

(See the section below on the social history of family relations and human development.)

Other Disciplines

A life-span orientation to behavioral change also emerged from the work of historians on family dynamics during industrialization in Europe and North America. Vinovskis (1977) notes that multidisciplinary interest in the historical family altered the traditional methods and conceptual models that historians used. For example, to recreate the structure and processes of family life they began to use quantitative data from censuses and vital records. Historical demographers such as Louis Henry (1956) and E. Wrigley (1977) used parish records, while Peter Laslett (1972) at Cambridge University reconstructed preindustrial family typologies from manuscript censuses. The latter scholarship revealed that conventional thinking about the large, extended family of preindustrial times was mistaken. In fact, family size was small, nuclear in form, and rather constant over time and cultures. But this work was later criticized by Lutz Berkner (1975) and others for its static approach to the family--in other words, for its lack of a dynamic, life-course perspective. Research into the historical family also drew from family sociologists and the concept of the family cycle (Glick, 1947; Hill and Mattessich, 1979). This work has attempted to understand the changing role of the family or household aggregate and the penetration of historical events and secular changes. In effect, the new behavioral approach of social historians has come to be a study of the dialectical relationships among individual time, social time, and historical time (e.g., Hareven, 1977; a similar set of distinctions about the various tempos of individuals' lives has been made by nonhistorians, e.g., Neugarten and Hagestad, 1976; Riegel, 1979).

Economists have a somewhat longer history of interest than historians in lifelong behavioral changes, although their contributions to an emerging multidisciplinary conception of human development and social change during the last decade have been slight. This lack of influence is rather surprising, since the ideas that individuals are human agents of production and that investments in human capital (e.g., skills and abilities) are components of production are old ones (Marshall, 1948; see Rosen, 1977, for a review of recent empirical research). A systematic statement of human capital theory by Becker (1964) prompted economists to develop theories of "permanent" earnings and to analyze lifetime decisions about work and leisure as well as longitudinal trajectories of earnings. For labor economists, this life-span approach was as revolutionary as were earlier developments in lifetime consumption decisions (Rosen, 1977:4; see Stigler, 1954, for a review of the history of economic research on consumption). Becker's (1965) seminal writings on the allocation of time to home production as well as to work in the conventional (paid) economy formed the core of the "new home economics," an approach that has led to the reorganization of university-based schools of home economics and increases in their faculties of behavioral scientists trained in economics and child psychology. The empirical work in life-cycle economics over the last decade has demonstrated a decidedly demographic character; for example, it has been applied to the analysis of fertility behaviors in connection with labor force participation (e.g., Easterlin, 1980). Some economic writings on social security also show evidence of a life-span orientation (e.g., Heckman, 1974; Modigliani, 1966). This work casts the economic decisions of the parental generation (e.g., investments in their own retirement security through time at work and savings and investments of time in developing the human capital of their children) in terms of the expected behaviors of children when the latter are of working age and the former are in the postproductive years.

Perhaps the lesser role of economics in influencing a life-span approach to behavior reflects the field's narrow focus on the optimizing, rational decision maker and the relatively minor impact of psychological economics (Katona, 1975) within the discipline. (One exception to this observation is the 14-year study by James Morgan and associates--e.g., G. J. Duncan and J. N. Morgan (1980)--of individual and family economic behaviors. These studies are rich in the behavioral and psychological data from which life-span analyses can be executed.) Moreover, the human capital orientation in labor economics takes a view of human development that contrasts with a central tenet of the life-span approach. Human capital theorists see the potentials for upgrading skills or changing the competitive qualifications of workers in rather determinate, age-graded terms. That is, workers (and their potential and actual employers) have a finite range of years (ages) in which to invest in improving or altering their stocks of human capital. The range is limited according to the theoretical assumption--now being challenged by life-span research in other disciplines--that capacities and interests to learn and to contribute productively to the economy are greatest in the young and deteriorate after middle age. If the corpus of emerging life-span research is assimilated by economics over the next decade, one should see some reorientation of both theory and research within the human capital framework. At the same time, the quantitative formalism of this economic approach holds promising benefits for life-span research in other disciplines, as illustrated, for example, by the influence of econometrics on life-span approaches to the study of social mobility and inequality in sociology (see the section on social inequality and stratification below).

Anthropological research on culture and personality played a major historical role from the 1920s to the 1940s in sharpening the concept of socialization as used by sociologists and anthropologists (Clausen, 1968). These studies were rooted in the influence of Franz Boas, primarily through his students, Ruth Benedict and Margaret Mead. Benedict's (1938) paper on continuities and discontinuities in cultural conditioning conveyed a life-span theme in characterizing the synchronies and asynchronies between biological changes in individuals over the life-cycle and the demands of culturally scheduled shifts in roles, responsibilities, and expected abilities. Edward Sapir (1934) was among the first anthropologists to emphasize the reciprocal relationship between personality and culture--to view personality as a transducer of cultural influences rather than as the passive expression of them. This view contrasted with the major theme of cultural continuity and the emergence of modal personalities within cultural contexts that mirrored the neo-Freudian psychoanalytic influence of Abram Kardiner (1939), even though the latter introduced an explicitly intergenerational linkage into the relationship between personality and the social system.

Culturally defined stages of the life-cycle have been used by ethnographers to organize and describe cultures. And a subdiscipline of anthropological gerontology has emerged since World War II in which the roles, statuses, and treatment of old people is the focus in both cross-cultural comparisons (e.g., Cowgill and Holmes, 1972; Simmons, 1945) and domestic studies (e.g., Clark and Anderson, 1967; Keith, 1979; Myerhoff and Simic, 1978). This body of work has served an important debunking function, showing how North American conceptions of the behaviors of the elderly may be culturally specific and not an inevitable feature of biological aging. An excellent example is the challenge to Cumming and Henry's (1961) disengagement hypothesis. Psychological and social withdrawal by the elderly--the lack of vitality, intellectual activity, and independence--are not universal behaviors, even among all North American ethnic groups (e.g., Cool, 1980; Kiefer, 1974; Vatuk, 1980).

As valuable as this anthropological work has been as a corrective on ethnocentric perspectives and as a sharpening influence on conceptual thinking, until recently it has not manifested an explicit life-span orientation. However, anthropological research on age as a basis of social organization is now under way. For example, culturally defined markers such as puberty, marriage of the first son, or death of a parent are being studied for their use as scheduling signals for transitions from one stage of life to the next--(e.g., Foner and Kertzer, 1978; Stewart, 1977). Other work focuses on individual differences in the course of aging and human development within a cultural setting (e.g., LeVine, 1978) and on the capacity of older persons to create cultural norms for age-related behaviors through innovative formation of senior citizens communities (e.g., Rosow, 1976).

LIFE-SPAN APPROACHES IN RESEARCH

It is premature to predict whether a life-span approach to the study of development and aging will ultimately blossom into a new discipline with coherent theory and special methods. One early (but not definitive) indication is whether applications of the thematic orientation produce cumulative social science and prompt the development of paradigm shifts within existing disciplines.

Since 1970 there have been promising signs for the vitality of the life-span perspective applied within the existing disciplines of sociology, psychology, and social history. The following section illustrates the usefulness of the life-span approach in three research arenas, in which it has reoriented concepts and methods, led to cumulative science, or prompted productive confrontations over metatheoretical differences. The sociology of life chances, psychometric intelligence, and the social history of family relations and human development each illustrate the application or explication of some basic life-span themes and propositions within research programs of three disciplines.

Social Inequality and Stratification: The Sociology of Life Chances

Over the last two decades, progress toward cumulative social science has been greater in the subdiscipline of social stratification and mobility research than in any other field of sociology. In recognizing this progress, there can be little dispute about the significance of Blau and Duncan's (1967) monograph, The American Occupational Structure, for sociological theory and research during this period. This work and the related writings of Duncan recast the empirical study of social mobility inquiries about the intergenerational and intragenerational processes of socioeconomic stratification.

Blau and Duncan provided a rudimentary life-span framework--the socioeconomic life cycle--for cumulative studies that extended and elaborated the descriptive features of stratification as a dynamic process of generational and cohort replacement in a society over time. This framework helped to organize and focus discussion about questions of inequality and the transmission of differential opportunities from generation to generation. It provided a focus for the discussions of public policy about poverty and human rights that prevailed during the 1960s and early 1970s as well as for debates between academic scholars. More generally, the work associated with Duncan and The American Occupational Structure became an exemplar for the design and analysis of national studies of mobility and inequality. The greatest impact of this program of research on the discipline may have been through its introduction of an approach to causal modeling of hypothesized relationships (e.g., social mobility) that could be applied to other areas as well.

Blau and Duncan cast the study of social mobility as the study of the process of stratification. Following Sorokin (1927), they conceived of mobility as a process of social metabolism whereby the inequalities that characterize society in one generation are reproduced, in whole or in part, in the next. By studying intergenerational mechanisms of socioeconomic transfer and factors that mitigate the effects of these mechanisms in people's lives, they investigated societal changes in the dispersion of socioeconomic statuses through the succession of generations. (Note the parallels with the age stratification model of Riley et al., 1972.)

Duncan's schema of the socioeconomic life cycle expressed this process of stratification in terms of the experiences of a hypothetical birth cohort (O. D. Duncan, 1967). It characterized inequalities in the cohort at birth by the socioeconomic statuses, genetic endowments, cultural and racial features, and related factors across households and communities. These inequalities of social background were taken as the antecedents of educational differences, which in turn were taken as the antecedents of differences in the occupational and economic statuses of the cohort. By studying differences in hierarchical standing across the successive stages or phases of the cohort life cycle, Blau and Duncan portrayed the pattern of social mobility over the life-span. Their framework permitted them to examine, for example, to what extent years of school attainment across individuals reorganized the patterns of socioeconomic inequality ascribed by social background. By comparing and contrasting this process of stratification in the experiences of successive cohorts, they were able to assess changes in inequality in society that were associated with changes in the relationships among social background, schooling, and occupational careers.

Path Analysis and Structural Equation Models

The impact of both this definition of social stratification and the framework of its study might not have been so pervasive or long-lasting without Duncan's introduction of path analysis as a statistical tool for sociological research (O. D. Duncan, 1966). Indeed, neither the conceptual point of view embodied in the socioeconomic life cycle nor path analysis itself was the discovery of Blau or Duncan. But the conjunction of the two was a powerful combination that both added to the potential of the Duncan-Blau approach to stratification and illustrated how sociologists might represent and study causal processes generally.

Path analysis, developed by the population geneticist Sewall Wright, enabled Duncan to partition the statistical correlations among the constituent phases of the socioeconomic life cycle (i.e., the relation between two instances of interindividual differences) into the quantifiable paths of direct and indirect influence between (hypothetically) antecedent and consequent events. For example, the correlation of social background and adult socioeconomic status (e.g., as indexed by parental and adult occupational prestige scores) could be decomposed algebraically into a precise statistical estimate of the direct effect of background on socioeconomic status and the indirect effect of background through schooling. In order to use this statistical method the analyst was forced to be explicit about the hypothetical model to be estimated: that is, to specify all direct and indirect relationships and to examine the variance left unexplained in each variable by the causal system of alleged antecedents. So, for example, Blau and Duncan could analyze the mobility-inducing effects of formal schooling that were independent of inequalities of social background--i.e., inequalities of background that were transmitted through schooling and converted into inequalities in the cohort's occupational attainments in adulthood.

The analytical power that path analysis provided for stratification research was twofold. First, its requirement for a precisely specified model and its capacity to provide statistical estimates of the model's credibility helped to formalize and make concrete theoretical discussions of processes of mobility. Analysts could visualize and critique each other's work with far more specificity about the entire system of relationships being considered and/or excluded. This facility increased the frequency of cross-disciplinary citation, especially between economists and sociologists nominally at work on the same topic; it also increased the rigor with which theoretical disagreements could be pursued. Second, path analysis led to rapid accumulation of descriptive findings and to a deeper understanding of the process of stratification. Because the technique was based on statistical (Pearsonian) correlations, analysts could synthesize complex path models from fragments of data across several independent inquiries, subject to the constraints of population and sampling comparabilities. This strategy of incremental model building is illustrated by O. D. Duncan et al. (1972), who elaborated and extended the basic five-variable model underlying the analysis of The American Occupational Structure. They introduced cognitive and motivational variables that were thought to intervene between social background and scholastic attainment, examined the potential of schools to affect the distribution of achievement apart from the personal and background qualities of students, and investigated the role of selected life-cycle events in adulthood in altering the pattern of socioeconomic careers.

Two other instances of the integrating effect of path analysis (structual equation methods) and the life-span approach appear in the writings of William H. Sewell and Melvin Kohn and their colleagues. Sewell has followed a longitudinal panel of Wisconsin high school seniors for over 20 years. The richness of the data that Sewell and associates have collected on social background, schooling, work histories, and life events from this cohort suits the application of structural equation methods. The Wisconsin status attainment model (e.g., Sewell et al., 1969; Sewell and Hauser, 1975)--a sociopsychological conception of social stratification--in some sense anticipated the Blau-Duncan framework for the socioeconomic life cycle. But by the mid-1970s a rapidly expanding literature had appeared in which analysts at Wisconsin and elsewhere both elaborated and replicated the quantitative statistical models of social (institutional), psychological (individual), and social psychological (interpersonal) factors in educational, occupational, and economic achievement (see Sewell and Hauser, 1980, for a comprehensive summary).

Melvin Kohn and associates at the National Institute of Mental Health have combined the traditions of this stratification research with pychological studies of mental health (Kohn, 1969; Kohn and Schooler, 1973, 1978). By following a national sample of adult male workers for roughly a decade, Kohn has studied how the requirements and organization of work and job changes influence workers' values, the goals they have in rearing their children, and even their cognitive or intellectual styles. This work also has adopted the quantitative statistical models of stratification research to explicate a life-span conception of the interplay between work histories and personality. For example, Kohn and his colleagues demonstrate the reciprocal relationship between changes in the demands of successive jobs (e.g., whether they demanded self-direction or were highly supervised; whether they called for the handling of complex novel circumstances or were substantially routinized) and changes in the intellectual capacities of workers in different occupational trajectories or job sequences. While it was true that men with greater potential at the outset were recruited more frequently into jobs requiring greater intellectual flexibility, it was just as likely that they were socialized by the job irrespective of personality or sociological factors at the beginning of the study. By implication, personality change (i.e., cognitive capacities) after childhood and adolescence can be observed readily through the study of successive role contexts that organize the tempo and content of adult behavior.

Life-Span Themes Emerging in Stratification Research

Sociologists who study patterns of inequality and social mobility in American society, as well as those who reflect on their implications for public policy, have begun to recognize that some of the same themes that characterize human development in life-span perspective also apply to the socioeconomic life cycle. In constructing this integration, the model of age stratification of Riley et al. (1972) (Riley, 1976) has supplied a highly useful conceptual schema. If one thinks of the process of stratification as consisting of lifelong trajectories of achievement behaviors, then growing evidence indicates that the developmental course of achievement (1) is responsive to many causes, (2) proceeds in many directions, (3) is both continuous and discontinuous, (4) entails greater interindividual differences as it unfolds, and (5) varies across the experiences of successive birth cohorts. In addition, this complex pattern of achievement throughout the life course is generated by age-graded events, cohort-forming events, and idiosyncratic events in each individual's life. As such, discontinuities in achievement from one phase of life to the next are to be as expected as continuities, and patterns of continuity and discontinuity in achievement are historically variable. Research evidence amplifies these points and illustrates an empirical base of life-span propositions and themes.

Some of the strongest continuities in achievement are intergenerational, for example, correlations between the parental and filial generations in performances on standardized IQ tests. Sociologists conceive of IQ as the measured ability to do schoolwork, and IQ score is, among other things, a performance on an achievement task. Interindividual differences in IQ tend to plateau between ages 8 and 10; thereafter, inequality (but not necessarily plasticity) in this form of scholastic achievement remains very constant through adolescence. Whether it continues to do so after adolescence is not well established (Schaie, 1979). Neither are the reasons for this developmental pattern well understood; it may reflect measurement error, lagged genetic effects, age-graded school environments, or all of these. The parent-child or intergenerational correlation of IQ is roughly 0.5, but the intergenerational discontinuity in this form of achievement is large. For example, a linear combination of social background characteristics--social class, race, region of residence--accounts for less than half of the variation among children's IQ scores (Featherman, 1980).

Other scholastic achievements (e.g., grades in courses, grade point average, teacher evaluations, years of school completed) and occupational attainments (e.g., earnings) are less connected to the similar intergenerational social achievements than is IQ. The life-cycle pattern is one of greater attenuation of discontinuity as the filial generation ages. In addition, secular trends or social change across the experiences of successive birth cohorts of Americans (and perhaps elsewhere) appear to be weakening these linkages even further. Take, for example, the length of formal schooling. The major predictor of length of schooling is not parental social class, but son's or daughter's IQ (Featherman, 1980). Educational aspirations, significant others' school plans, IQ, and grade point average explain about 70 percent of the educational differences among individuals. Interestingly, the net effects of parental characteristics are effectively zero in these analyses. And across successive birth cohorts of American men, these associations of social background (e.g., race, class, farm orgins) are getting weaker (Featherman and Hauser, 1978). At least this is so for education through grade 12. This declining persistence of achievement from generation to generation stems in large part from secular reduction in educational differences at the elementary and high school levels. Perhaps this reflects a legacy of industrialization and child labor laws. In any case, historical change has altered the intergenerational pattern of continuity and change in scholastic achievement.

Turning to occupations and careers, one finds, at least for the United States, secular declines in the predictability of achievement from the attainments of parents and associated aspects of social background. But unlike the situation for length of schooling, this trend toward greater discontinuity in socioeconomic achievement is not connected to overall reductions in occupational inequality--in the distribution of the prestige of jobs on some scale of social standing or income. Rather, it seems to arise from the substitution of formal education for social class or background as the means of access to better jobs or careers. Obversely, the continuity between achievement in school and at work has improved, especially the relative economic value of higher education (Featherman and Hauser, 1978).

Explaining the overall trend toward greater continuity between the scholastic achievements of youths and the occupational and economic attainments of adults is problematic. Whether it reflects greater valuation of higher education in postindustrial society, the effects of "credentialism" in the allocation of workers to slots in the economy, or both, is not known from available research (Featherman, 1980; Featherman and Hauser, 1978).

There are some major exceptions to this historical trend. One involves black workers, for whom intergenerational continuities of achievement--modest though they are for Americans overall--are just now beginning to approximate the pattern that has been typical in white families (Featherman and Hauser, 1978). Another exception involves white men who were under age 35 in the mid-1970s. For these men the economic value of higher education seemed to have fallen as they took their first jobs. On the basis of recent analyses, it is possible to interpret such a cohort pattern as a temporary aberration that stemmed from a unique confluence of demographic, economic, and historical events (unprecendented cohort size, downturn in federal expenditures for research and development, and the effects of the Vietnam military draft on school attendance patterns; see Featherman and Hauser, 1978). But the important point for a consideration of lifetime continuities in achievement and their vulnerability to historical and social change is that this unique confluence may have cost the college class of 1974 in the United States about 10 percent of its lifetime earnings (Welch, 1979).

The normative features of the age-graded life cycle are linked to achievement patterns, too. For any given birth cohort there is a statistically normative age profile to the entrance into and exit from the family of origin, school, work, and the family of procreation (Hogan, 1978). Whether such age-graded behaviors are socially as well as statistically normative--subject to positive/negative sanctioning--has yet to be firmly established. But there are consequences for achievement that ensue from deviations from the normative order and pace of life-cycle events (e.g., Hogan, 1980). These effects are easiest to illustrate in contrasting the connections between jobs and schooling for American women with those of men. Secular trends in the American female life cycle during the 1970s have markedly changed the labor force participation of women in the prime marital and childbearing ages--25 to 34. Whereas two decades ago only one-third of such women worked, today about 55 percent do. Of women in that age bracket 75 percent who do not have children are employed. Over 90 percent of the men in these ages are in the labor force, however, and they tend to work with fewer interruptions and more frequently in full-time jobs than women. Thus, despite recent shifts that have rendered the female life cycle more like that of men, the family cycle is still more integrally related to the socioeconomic life cycle of women than of men (Van Dusen and Sheldon, 1976).

What are the consequences for achievement? Relative to men, working women tend to experience more downward social mobility as they raise their families and find their own careers. They acquire less job experience at each age than men, making job-to-job moves less predictable and less conditional on job characteristics than they are for men. There is less continuity of occupational achievement for women (Dunton and Featherman, forthcoming). In addition, the connection of schooling to successive jobs differs. For men the direct influence of formal schooling on jobs is greatest at career beginnings (i.e., first job). It declines thereafter as experience and on-the-job training become more important for subsequent career moves. For women, however, formal schooling retains importance as the major access to subsequent jobs, as women are forced to renegotiate for new jobs on the basis of their school credentials or formal training rather than on a stream of cumulative experience. Overall, however, the net effect is for less continuity between achievement in school and in work for women as a function of deviant age-grade patterns in their socioeconomic life cycle (Sewell et al., 1980).

In this brief summary of research on the process of stratification one sees the imprint of a life-span orientation. Achievement behaviors across the life cycle take place in a sequence of institutional contexts--the home, the school, the workplace, the economy. The age-graded features of this sequence give rise to the socioeconomic life cycle as one aspect of the general life course. Social changes in the connections between these institutions alter the pattern of continuity and change in human development, as do factors that independently may affect the sequence and pace of life-cycle transitions. In the case of achievement behaviors, especially for men, the drift of change in social institutions over the last several decades has been to reduce the possiblity of continuity in the developmental differences among individuals as they age.

PSYCHOMETRIC INTELLIGENCE: COMPETING INTERPRETATIONS ABOUT DEVELOPMENTAL CHANGE IN ADULTHOOD AND OLD AGE

Developmental Aspects of Intelligence

Within psychology, theoretical dissensus has surrounded the study of cognitive development in adults and particularly the aged. For instance, the putatively minor developmental changes in adulthood and universal senescent declines in old age clash with evidence and points of view drawn from the emerging life-span paradigm (compare Horn and Donaldson, 1980, and Willis and Baltes, 1980). Despite such theoretical dissensus, psychologists studying psychometric intelligence have come to agree that global measures such as single IQ scores can no longer be used productively in developmental research with either children or adults (McCall, 1979; Schaie, 1979). Instead, in the last decade an approach to the testing of a variety of primary mental abilities throughout the life-cycle has developed. Each of these abilities (e.g., spatial relations, visualization, verbal comprehension, symbol manipulation) is identified by one or more specific tests and can be scored individually. In turn, these primary mental abilities--ranging from 10 to 120 in number--have been found to cluster into a small set of higher-order factors or latent mental attributes that are reflected in the primary abilities. According to one widely accepted model, the most central of these latent attributes are fluid and crystallized intelligence (Horn and Cattell, 1967). Crystallized intelligence refers to the universe of abilities embodied in the symbolic culture of a society; it comprises knowledge and mental skills that a community deems valuable and essential for its maintenance, which are instilled through childrearing and adult socialization. In broad terms, crystallized intelligence can be seen in the ability to decode a written or oral message, to identify its main ideas, and to retain them for later use; it underlies the capacity to cope with social situations according to the conventional mores of a community; it can be seen in the ability to balance a checkbook or to fill out an IRS Form 1040 each April.

In contrast to crystallized intelligence, fluid intelligence deals with reasoning about novel or unfamiliar material. Because fluid intelligence is seen as the developmental basis for all intelligence, it involves many of the same capacities as crystallized intelligence--the abilities to abstract, to solve problems logically, and to cope intelligently with everyday life. Yet it is distinguished by a unique set of manifestations that are not easily taught by parents or the schools or learned on the job. Rather they are thought to arise through incidental or casual learning and through other influences that affect the physiology and neural processes associated with intellectual development (Horn and Donaldson, 1980:461). Fluid intelligence involves capacities that enable a person to reason with abstract symbols, to invent and use alternative optional or unconventional classifications or problem-solving strategies, or to visualize novel or hypothetical events.

Both fluid and crystallized intelligence themselves are organized hierarchically as reflections of general intelligence, which comprises the concepts that global IQ scores index. What developmental research missed as it studied the age trajectories of global IQ scores (e.g., Bloom, 1964) was that fluid and crystallized intelligence apparently have different developmental profiles. For example, Horn and Donaldson (1980; see also Schaie, 1979) review a corpus of age-related psychometric research of the last decade that concurs with the view that fluid intelligence declines after ages 25 to 30, whereas the abilities associated with crystallized intelligence suffer no decline or improve during adulthood (i.e., up to the retirement years). While the mental faculties tied to memory also decline as a function of chronological age and do so concurrently with fluid intelligence, there is no apparent causal connection between the two losses in the research they report. The explanation that Horn and Donaldson provide is a multicausal, highly speculative one, inasmuch as the research base that would be required to sort out the relationships has not yet been assembled. They suggest that adult life provides opportunities for practicing and sharpening the primary abilities subsumed by crystallized intelligence. The research of Melvin Kohn and associates (Kohn and Schooler, 1978) linking the substantive complexity of men's work, their sequences of occupations and job-related tasks, and their profiles of intellectual flexibility illustrates one such long-term arena for learning and practice. By contrast, there may be fewer and fewer contexts within which the incidental learning of fluid intelligence takes place in adulthood. And the accumulation of brain damage and neural dysfunction as a function of age (the sheer passage of time and increased exposure to the risk of injury) seem to have a greater effect on fluid than on crystallized intelligence. Thus, for example, between the ages of 20 and 60 there is about a 12 percent loss in brain weight, a decrease (up to 50 percent in some areas of the brain) in total number of neurons, an increase in neurofibrillary tangles and plaques, an increase in the width of brain fissures, and so on. Horn and Donaldson conclude: "The empirical evidence indicates that the abilities of G_f (fluid intelligence) are more permanently affected by loss of brain tissue than are the abilities of G_c, but it is by no means clear why this should be true" (Horn and Donaldson, 1980:480).

One reason may be that the knowledge structure of crystallized intelligence is overdetermined--that is, it seems to be based on neural structures that contain a higher order of redundancy than those of fluid intelligence. Loss of brain tissue therefore may not be as crucial to the maintenance of full capacity. Horn and Donaldson's review of neural physiological research suggests a basis for this differential redundancy. It appears that crystallized intelligence may depend on biochemical structures, or biochemical alterations of neuronal synapses, that seem to be the mechanism whereby information is stored diffusely throughout the brain. Fluid intelligence, by contrast, appears to depend on electrical networks of firing neurons. Loss of even a small number of neurons would impair the action of an entire network, whereas a similar loss would have less effect on the diffuse biochemical structure. This speculation, tentative as it is, illustrates the close interdependence that has evolved over the last decade between the bioneural and the psychological sciences. It suggests that cross-disciplinary exchange among biologists, psychologists, sociologists, and other behavioral scientists will be essential in moving the life-span orientation to intellectual development into concrete and theoretically focused research programs.

Horn and Donaldson's review concludes that age-related declines in the capacities of fluid intelligence are linked to deterioration in the ability to encode (to organize rather than to retrieve) information and to maintain close attention to the details of complex problems or to conceptualize nonstandard relationships. Presumably these age-related, biologically based declines in intellectual skill lie at the root of naturalistic observations by Lehman (1964) and others (reported in Horn and Donaldson, 1980:494). For example, poets are said to peak at ages 25-29; psychologists, at ages 35-39; the rate of output of chemists is highest at ages 30-34 and drops by 40-44; peak years for largest annual earnings in most fields are ages 50-55. (Lehman's unproven estimates have been challenged; see Riley and Foner, 1968.)

The Life-Span Critique of Developmental Conclusions

Horn and Donaldson reflect an orientation to intellectual development that spans the adult years into old age and highlights the multidimensional, multidirectional, multicausal features of development. This orientation has been challenged by Schaie, Baltes, and others as anachronistic in light of recent life-span research (P. B. Baltes and K. W. Schaie, 1976; Schaie and Baltes, 1977; Willis and Baltes, 1980). Furthermore, life-span research suggests that the assumption of a biologically inexorable deterioration of fluid intelligence and related capacities in all individuals in all cohorts does not square with new evidence about the effects of intervention among those over age 65.

The critique refers to the tendency of Horn and Donaldson to view intellectual development in terms of the normative biological growth model that guided child development and gerontological research throughout most of this century. It finds fault with the emphasis on normative, developmental functions that fail to give attention to the variability of abilities and performances. Such variabilities around the Horn-Donaldson norms manifest not only changes within the individual and his or her context over time, but also biohistorical changes across samples of different birth cohorts and sociocultural differences among individuals within a given cohort.

Plasticity and Variability in Intellectual Functioning

Schaie, Baltes, and their colleagues at the frontier of life-span research on intellectual development emphasize the plasticity of intellectual functioning over a person's life; that is, the apparent capacity for marked increases or improvements as well as for deteriorations in mental abilities and performances in terms of both fluid and crystallized intelligence until death. They also emphasize that not all individuals age intellectually in the same way; variability in fluid and crystallized intelligence across the age profiles of individuals is substantial, especially across cohorts and across persons with different life histories. Evidence for these conclusions is tentative (e.g., Horn and Donaldson, 1980), but it is becoming firmer as longitudinal cohort studies and experiments with interventions are conducted with these ideas as guiding hypotheses.

For example, Schaie (1979) has studied a large number of persons between the ages of 24 and 80. Dividing the sample into seven birth cohorts, he assessed the change in primary mental abilities at three points over a 14-year period (1956, 1963, and 1970). He was able to analyze age-related (ontogenetic) changes by comparing the change in scores for persons at comparable ages (e.g., 35 to 41; 42 to 49); departures from a common ontogenetic pattern across the 7 groups were interpreted as cohort differences. In general, Schaie found that cohort variation was greater than the magnitudes of change in ability that could be attributed to aging within cohorts. Cohort differences were not uniform across the component primary abilities associated with fluid and crystallized intelligence. This work challenges the assumption of a universal normative pattern of intellectual performance across the life-span; it offers contradictory evidence for the assumption of inevitable and uniform declines in fluid intelligence at advanced ages. (Obversely, it contradicts the generalization that all persons necessarily enjoy stability or improvement in their capacities for manifesting crystallized intelligence.)

There are, of course, many reasons why successive birth cohorts might display different capacities for growth and decline across the various components of fluid and crystallized intelligence. Uhlenberg (1979) has described the massive demographic changes under way in the characteristics of the elderly population in the United States. It is not at all unlikely, given the overall improvements in levels of education, health, and economic security, that the pattern of future research such as that of Schaie will show smaller declines in mental ability at every age in successively more recent birth cohorts. Opportunities for the learning and practice of abilities associated with crystallized as well as fluid intelligence have improved and may continue to do so. Improved health care and protection from hazards and injuries may foster greater neural capacity at advanced ages. On the other hand, the course of socioevolutionary change is not inevitable, and the neurons and biochemistry of the brain may be subject to a genetic, evolutionary program for eventual dysfunction and death (e.g., Strehler, 1977). Despite the inevitability of death and dysfunction at some chronological age, the situational context is varied for persons within cohorts and across them as well.

Another program of research on psychometric intelligence, guided by Baltes and colleagues, offers suggestive evidence about the plasticity of fluid ability in the elderly (e.g., P. B. Baltes and S. L. Willis, 1981; Willis and Baltes, 1980; see also Denney, 1979; Labouvie-Vief, 1976; Sterns and Sanders, 1980). Baltes and his colleagues have addressed themselves to the underlying potential or reserve for performing various intellectual tasks across the life cycle. They reason that mental abilities must be differentiated from measured performances of tasks that call for the application of ability. On any single occasion, factors such as motivation, fatigue, stress, and the like may alter performance and introduce situational "error" into the estimation of an individual's ability. Similarly, across the life-span, situational contexts influence performance and add to the interindividual and intraindividual variability at intellectual tasks. Thus, what the analyst observes is some interaction between the "true" or latent ability and the environmental context. But what about ability in some optimal environment, one that is structured to reveal the latent potential or reserve of ability? In a series of intervention or optimization demonstrations with persons between ages 60 and 80, the Baltes-Willis group has found substantial reserve for improvement in performances at tasks that tap fluid intelligence. Not only do old persons do better at the laboratory tasks, but also the experimental treatments seem to encourage generalization to other tasks as well (see P. B. Baltes and M. M. Baltes, 1980, and Labouvie-Vief, 1976, for reviews).

Baltes and colleagues have reasoned that psychometric tests are performances that reflect both competence and situational influences such as fatigue, motivation, and interest. Insofar as some primary abilities associated with fluid intelligence decline in some older persons, might these changes be tied to the situational factors rather than to ability per se? In a series of experiments, the Baltes group has attempted to optimize the performances of persons aged 60 to 80 who were drawn from a university community.

One group was tested at repeated intervals in order to give implicit familiarity with the testing situation and the tests themselves. Another group was given explicit and specific training in eight one-hour sessions at the problem-solving skills that the tests were designed to measure. Still another group was tested only at the very end of the experimental series and represented a posttest control. With the exception of this last group, the others were tested following the training period at one-week, one-month, and six-month intervals. Sheer familiarity with test taking seemed to improve the scores in the first group. But marked improvements in performances were apparent in the group given explicit training at the rules and logic of tests measuring primary abilities associated with fluid intelligence. Even six months after the actual training had ended, the participants continued to improve at a rate that exceeded the gains in the "familiarity" group. And those with explicit training also were able to generalize their new skills to tests of fluid intelligence that had not been the focus of specific training. In related work, tests of response speed also showed the latent capacity for small modification.

These studies have suggested that losses in neural functioning that the elderly suffer may not always impair intellectual performance. Practice at new tasks (or at old ones that have ceased to be salient, such as test taking for a 70-year-old), motivation, reinforcement, and focused attention are some of the situational factors that have the capacity to mediate the significance of biological changes for the behaviors of the aged. The Baltes group has implied that the elderly have traditionally lived in a context of ill-defined social roles--or perhaps more correctly, a roleless phase of life (Rosow, 1976). Lacking practice or opportunities to learn and sharpen skills, their abilities fall into disuse and deteriorate. (The same argument often is made by human capital economists in interpreting the lesser economic returns to investments in education or prior job experience by middle-aged women vis-à-vis men with more continuous work histories; e.g., Polacheck, 1979.) This interpretation implies that intellectual performance, and perhaps other abilities and capacities as well, are underlain by a latent reserve or potential that is only partially tapped by conventional social environments. Whereas children and adolescents manifest a greater proportion of this latent reserve, due to the orientations for achievement and personal development that are built into institutions such as the school and the economy, the reserves of the elderly are less fully utilized or revealed (e.g., P. B. Baltes and S. L. Willis, 1981).

Toward an Integrated Interpretation

This speculation may ultimately provide a basis for integrating the results of the optimization experiments with the work of Horn and others that posits eventual deterioration of function as a normative feature of age. The ontogenetic course of latent reserve may in fact peak in early or middle adulthood and decline thereafter because of genetic and other biological influences. But since this is a hypothetical and as yet unobserved limit of development, more age-comparative optimization research will be required to establish its factual basis. On the other hand, the manifest or actual reserve seems to be malleable in ways that call into question the inevitable and universal correspondence between its developmental course and the limits of latent reserves. In more optimal environments, the trajectory of manifest capacities may continue to rise into old age, long after the latent capacities have peaked.

Speculations of this character are prompting continuing research on the conditions of senescent declines in mental capacity and in independence and mastery behaviors among the elderly (e.g., M. Baltes and E. M. Barton, in press; Rodin, 1980). From this work have come a series of methodological insights that will continue to reorient developmental research. One obvious illustration is the use of optimization interventions that manipulate the situational contexts within which behaviors and abilities become manifest. Another is the development of new tests and instruments for the assessment of competencies across the full span of life. Schaie and other gerontologists have argued that psychometric instruments that were designed to measure differential abilities in children and adolescents are poorly constructed for use among persons in the later periods of life (Schaie, 1979). That is, the achievement-related contexts within which assessments of mastery and competence derive their originating purpose--i.e., to predict success in school and in the early work career--are largely without direct counterparts in the last third of life.

Schaie (1977-1978; see also Labouvie-Vief, 1980) has suggested a life-span theory of intellectual development in which the definition of intelligence changes according to the changes in developmental tasks throughout the successive phases of a typical life course. Fully recognizing that cohort and individual differences in life events may limit the usefulness of a normative approach to conceptual definitions, Schaie then suggests that new tests of primary abilities be constructed to tap these various dimensions. In addition, he and others have begun to explore the likelihood that the structure of psychometric intelligence also changes across the life-span. For example, the primary abilities that cluster into fluid and crystallized intelligence may undergo a transformation and realignment over time in response to both biological changes and sociohistorical ones. Results from early work seem to bear out this hypothesis and to imply that aging entails qualitative as well as quantitative change in intellectual ability (e.g., P. B. Baltes et al., 1980a). Application of structural equation modeling and the use of new computer algorithms for confirmatory factor analysis (e.g., Jöreskog and Sörbom, 1979) have aided this line of inquiry.

Life-span interpretations of intellectual development have animated conceptual and methodological discussions in psychology and fueled theoretical debates. While it is far too early to predict whether this dissensus and scholarly dialectic will catalyze a new behavioral science around life-span issues and methods, the last decade has witnessed a diffusion of new perspectives and methods into developmental psychology. Established concepts and modes of research design are being questioned, and there is a reaching out to related disciplines for help in addressing old questions in new ways. These studies and developments also have brought academic research on psychometric intelligence into closer relationship with practical or policy-related issues about aging and the elderly--issues like retirement, social security, independent living, and long-term care.

THE SOCIAL HISTORY OF FAMILY RELATIONS AND HUMAN DEVELOPMENT

Over the last decade, social historians, family sociologists, developmental psychologists, and demographers have begun to study the family through a life-span orientation (e.g., Hareven, 1978; Hill and Mattessich, 1979; Vinovskis, 1977). An interdisciplinary approach has become the explicit basis on which research problems about the family are defined, variables are selected, and analyses are designed (Elder, forthcoming). This trend has been facilitated by a willingness of the disciplines to expand the frameworks of analysis to include variables normally found outside their separate domains and by the accumulation of common data bases that permit integrated analysis (e.g., Hareven, 1978; Hershberg, 1981; Thernstrom, 1964). As a result sociological understandings of the contemporary American family are being transformed through revisions of stereotypes about the historical family; sociological and demographic insights into family process are revolutionizing historical research; and the psychology of ontogenetic change is adding a biobehavioral dimension to the analysis of family process. Perhaps because it includes so many of the multidimensional, multilevel issues in the analysis of the dialectic of individual and social change, the study of family relations and human development may provide the intellectual context for the eventual emergence of a new life-span discipline.

Reinterpreting the Historical Family

Elder (forthcoming) has observed that as little as a decade ago students of the American family were convinced of two generalizations about the historical trends in family life between the 19th and 20th centuries: that domestic households had become increasingly more nuclear (that is, they were composed of two parents and their children) and that they had lost many of their economic and developmental functions. Research in Britain by Laslett (1972), in Austria by Berkner (1975), and in North America by Thernstrom (1964), E. A. Wrigley (1972), and others is contraverting these assertions. This work is revolutionizing the concept of the historical family and, by extension, recasting our appreciation for continuities and changes in the contemporary family.

These radical shifts in understanding and in research approaches have been sparked by several developments. One is that manuscript censuses for the late 19th century have become available as research tools; this has enabled quantitative historians, demographers, sociologists, and others to synthesize the records of household members over successive enumerations and to link these to administrative and other secondary sources of information about employment, education, and income within the communities of residence. Primarily it has been efforts to synthesize processes of family dynamics with historical change within the framework of a life-course orientation that are enabling researchers to begin the difficult analytical tasks of rewriting history and of restating the status of contemporary family life within that historical pattern. Because these revisions are still in early stages and the data are still being assembled, only the general thrust of the work and preliminary findings can be reported.

Families in the 19th century are not more easily characterized than their counterparts today, our stereotypes notwithstanding. While on average the historical family may have been larger than most contemporary ones, household size and composition were highly variable. For example, 19th-century households in North America appear to have adjusted their size in relation to changing economic fortunes. In Canadian mercantile centers, young adults of working-class origin often spent a period outside their parental homes as lodgers in other households, as domestics, or as employees in firms at some distance from their families (Katz, 1975). In a sense they were part of the parental household economy, because they frequently shared some or all of their income with that unit and returned to it after a period away. Elsewhere the same families that dispersed employable members also took in boarders and lodgers as coresidents as needs arose (e.g., Modell and Hareven, 1973), either because of secular business cycles or because of life-cycle changes within the household, such as widowhood.

Thus families responded to industrialization and the urbanization of the 19th-century economy by adjusting their household economic bases in two ways that affected their size and composition: by expanding their sources of income and by limiting the demand on these sources within the household. At some times households were large and *extended*, at other times they were *stem*, and at still others they were *nuclear*. There was great variation over time and across households in the strategies that families used to respond to secular (historical) change and to life-cycle transitions and events of their constituent coresidents. In these respects, 19th-century families were no different from families today; see, for example, the only ongoing longitudinal study of American households, the Panel Study of Income Dynamics (e.g., G. J.

Duncan and J. N. Morgan, 1980). By comparison, contemporary households adjust their labor supply through the life-cycle and the secular pattern of women's labor force participation, family size (birth control), and child spacing. In the Morgan et al. studies, the chief factors that accounted for the economic status of households over time were those connected to shifts in their relative sizes and compositions: divorce and separation, additional children, and the separation of subhousehold units (G. J. Duncan and J. N. Morgan, 1976).

Dynamic and Behavioral Approaches

The major change in the study of the historical family has been the shift from an essentialy static, structural perspective to a dynamic, behavioral one (Elder, forthcoming; Hareven, 1977). Typological thinking about the preindustrial, industrial, and postindustrial family, families whose internal structure was thought to mirror faithfully the structural transformations of the surrounding economy and society (e.g., Smelser, 1959), has been replaced by approaches that see the family as a dynamic unit over the course of its life, changing its structural features in response to social change and to the life course or developmental trajectories of its members over time.

For example, Michael Anderson (1971) and Tamara Hareven (1981) use longitudinal historical data on the individuals within families and households to portray the active role of the family unit in the course of industrialization in both Britian and the American Northeast. During the early stages of industrialization, especially in the textile towns of the American Northeast, the economic survival of the family was well served by a collective strategy or family plan that sent women to work, withdrew children from school, or aided in the migration and job placement of kin as dictated by the changing fortunes of the family. This interplay of family time and industrial time (Hareven, 1975, 1977) gradually gave way, under conditions of rising affluence and declining family size, to a 20th-century pattern of individual life plans that could be pursued without jeopardizing the survival of others. During the 20th century, the earlier patterns of contingency were weakened between a person's transition into adulthood, including the assumption of independent economic roles and the making of a new family unit, and the obligation to be responsive to episodes of economic misfortunes in the parental household (Modell et al., 1976). Thus the conjunction of family time and industrial time was tranformed into one between individual time and industrial time. The tempo of the individual life course was organized by a new set of institutions outside the family that ordered the roles into and through which persons passed. Age-graded schools, occupational and industrial careers, and promotional schemes involving seniority are examples of the emergence of temporizing influences that not only affected the pace of an individual's life but also were generalized across individuals to form age-graded normative events as cohort experiences.

In this latter respect the behavioral approach to the 19th-century family as both the receiver of historical influences and an active agent in the course of historical change parallels the emergence of modern home economics as an orientation to the study of the contemporary family (e.g., Becker, 1965). Studies of investments in childrearing, the labor supply of mothers, and schooling (e.g., Kaestle and Vinovskis, 1979) in 19th-century America cast the family as a group of decision makers optimizing their utilities through production, reproduction, consumption, and resource allocation in a changing social and economic environment. What is striking about the new family history and the new home economics (which is really a version of human capital theory applied to the time allocations of production and consumption associated with domestic versus market labor decisions and childrearing) is that both theories have such underdeveloped conceptions of human ontogeny. Perhaps this weakness, or lesser theoretical development, is understandable in the reconstructions of 19th-century behavior, because of the absence of appropriate data. In the case of human capital theories of contemporary domestic economies, however, this conceptual shortfall is a challenge for the future. In this regard it seems essential that economists become a more central part of the multidisciplinary discussions of life-span behavioral processes. Longitudinal data on individuals within household aggregates, such as those being collected in the Panel Study of Income Dynamics by the economist James Morgan, are also indispensable to this endeavor.

Life-span behavioral approaches to the study of historical families as responding to and shaping the course of industrialization are reshaping current debates about the social functions of schooling and family in the political economy. New data from manuscript censuses and reconstructed life histories challenge the social criticism of radical economists and neo-Marxian sociologists that the schools in the 19th century were used by capitalists to control and shape working-class and immigrant behaviors into forms that were useful to entreprenuers (e.g., Bowles and Gintis, 1976; compare Kaestle and Vinovskis, 1980). By contrast, the picture of school enrollment and its relationship to the economic activities of families and the life courses of individuals is becoming much more complex. Studies show, for example, that school enrollment of older children was common prior to the middle and late 19th century and the spread of industrialization in America (e.g., Kaestle and Vinovskis, 1980). In addition, many of the modern attributes of the family (e.g., companionate social relationships) may also have been prevalent in the 19th century (Wrigley, 1977). Thus historical life-span research is revising the thinking about the modern family and its relation to the political economy, if only because social criticism of contemporary institutions is predicated on apparently inaccurate assumptions.

History and the Changing View of Modern Families

New thinking about the interconnection of social change, individual change, and change in family structure and process is challenging more than our stereotypes of the historical family. It is revolutionizing the study of the modern family as well (Elder, forthcoming). For decades sociologists and demographers have used the concept of the family cycle to describe and analyze regular changes in the social relationships and orientations among family members as a function of temporal shifts in family composition (Duvall, 1971; Glick, 1947, 1977; Glick and Parke, 1965; Loomis and Hamilton, 1936). That is, the family as a social aggregate was thought to assume certain universal behavioral features as a consequence of its structural properties and transformations of them. For example, Hill and Mattessich define family development as "the process of progressive structural differentiation and transformation over the family's history, . . . the active acquisition and selective discarding of roles by incumbents of family positions as they seek to meet the changing functional requisites for survival and as they adapt to recurring life stresses as a family system" (1980:174).

Historically, the family history to which Hill and Mattessich refer was conceived as a series of sequential, static, age-graded types of family structures--for example, marriage and the dyad, the birth of the first child and the triad, the youngest child's leaving home and the empty nest, retirement and the ultimate dissolution of the marriage through death of one spouse. Demographic regularities in age at marriage and in child spacing have provided the age-graded character of the family cycle, since the stages or phases of family life are predicated on markers of family time, such as the age of oldest or youngest child, the age at retirement, or the age at the death of a spouse.

Family researchers now recognize that the concept of the family cycle has been historical, static, culture-bound, and unduly focused on the effect of children on the parental relationship. Although the concept has aided the analysis of longitudinal change in behavior, it was based on typological thinking and on assumptions about the prevalence of marriage and the nuclear family and the durability of marriages throughout a lifetime. Glick and Norton (1977) project that among contemporary young marriages, 40 percent will end in divorce. Of those who become divorced, between three-quarters and five-sixths will remarry and remain in that relationship until the death of the spouse (Glick, 1977). Together with the greater stability of marriages in older birth cohorts, today's marital patterns attest to the preferableness of married life, for about 84 percent of all families in 1975 were husband-wife families (Glick, 1977). At the same time, the recycling of adults through marital relationships and the accumulation of children exposed to divorce and either long periods of single-parent family life or second or multiple families have increased markedly over recent decades. Only 67 percent of all children under 18 live with their own once-married parents (Glick and Norton, 1977). In this context the static typological model of the family life cycle has little scientific utility.

Hill and Mattesich's definition of family development is a heuristic effort to revise the study of the modern family that incorporates the elements of a life-span orientation and is flexible enough to apply across the historical experiences of different birth and marital cohorts (see Clausen, 1972; Spanier and Glick, no date). The newer approach views the family as a constellation of individual life courses in some mutually contingent relationship and in the context of evolving historical circumstances. The pertinence of research on family life during the industrial revolution for the conception of the modern family is that historical circumstances are themselves the outcome of the interplay of individual developmental processes, of family or collective responses to the historical moment, and of the opportunities and constraints of historical events and contemporary institutions.

Glen H. Elder, Jr.'s description of families in the Great Depression illustrates this interplay. Elder (1974, 1978, 1980) has conducted social psychological research on the reciprocal relationships between historical and personality changes using two studies of children born in the 1920s in the San Francisco Bay area as they have grown into adulthood. His work illustrates the cohort sequential method of longitudinal research, which contrasts the life courses and personalities of individuals in two birth cohorts as they age. In each cohort Elder explores the differential impacts of relative economic deprivation associated with the Great Depression and of mitigating influences associated with military service in World War II, subsequent career security, and other adult life events. One cohort, drawn from the city of Oakland, was comprised of people who were adolescents during the depths of the depression. A second cohort, from Berkeley, was comprised of people born later who spent their childhood in the postwar economic boom. Effects of sudden economic hardship and resulting family distress were more visible in the Berkeley cohort, for whom a greater portion of childhood was spent in hard times. The timing of this deprivation, relative to developmental age, placed the Berkeley children at greater risk of cumulative disadvantage than the Oakland children. Across the decades of longitudinal data, Elder observed men from households that suffered large economic losses--irrespective of social class--to voice concerns for security and to value financial conservatism. These attitudes and related behaviors were much less salient for men with stable work histories and marriages, demonstrating the moderating influences of proximate life events among men of equally deprived backgrounds. All men of either cohort were not equally affected by the depression. For example, some lived in families that lost relatively little; some had fathers whose sudden decline in earning power altered the pattern of parental dominance and the strength of the father as a role model for his son. Others came from households in which creative coping with distress and collective sharing of new responsibilities were sources of family solidarity. In each of these instances of differential loss and of family response, Elder found different manifestations of the depression--manifestations in anxieties and mental health and in values about conservatism versus risk-taking.

One instructive feature of Elder's long-term program of research from a life-span perspective is his demonstration of the importance of cumulative life history as a tool in the analysis of differential outcomes. He was able to document different manifestations of the depression in the preadult lives of the San Francisco Bay area residents and in the interaction between developmental age and the onset of economic hardship. In addition, Elder emphasized that the developmental consequences of the differential effects of the depression were even more varied in adulthood. For example, adolescents whose family's relations were heavily strained by the father's loss of substantial earning capacity and related esteem and whose mothers often assumed a position of dominance frequently suffered anxieties and doubts about personal competence. Yet if the son was able to move away from the parental household rather quickly, as was the case for many of the Oakland boys who were mobilized in World War II, long-term effects of these experiences in youth were offset by fresh starts in new settings. Others, who went to college, failed to evidence any career-related disadvantages, inasmuch as they were the most able to avail themselves of the expanding economic opportunities of the postwar boom and to establish the actuality of their competence.

In summary, Elder's continuing longitudinal research on these two cohorts reflects many of the themes and propositions of the life-span orientation. For example, developmental research must be historical and situational insofar as historical events precipitate change in the course of lives, both between and within cohorts. And developing individuals are agents as well as receivers of historical change. (By implication, Elder's work suggests that there may be only a limited set of generalizations that are ahistorical, i.e., as true in 1990 as in 1980, that behavioral scientists can make. This may differentiate social science from other sciences.) The chief illustrative value of Elder's work may lie in its description of the ways in which a single historical event interacts with the circumstances of people's pasts and futures to increase the likelihood of both change across the course of life and of individualizing of the life trajectories of adults from seemingly similar social and historical origins.

Usefulness of the Family as a Unit of Analysis

Viewing families as coresiding individuals complicates the study of the family, for it forces the analyst to see a family unit as a potentially unique entity. Each individual is at his or her point in personal developmental time, the significance of which is cast in terms of a cumulative life history. The aggregate coresidential unit can change in time because of dissolutions of marriages, remarriages, or other compositional changes that imply a dynamic situational context for those individual life courses and their combinatory outcomes. Then, too, there is the impact of sociohistorical change that may become manifest in unique ways.

Sociological and economic research on cohort marital fertility and female labor force participation supports this perspective (e.g., Easterlin, 1980), as does Alice Rossi's (1980) examination of how the hormonal and physical changes (or differences in the degrees of biological change) in middle-aged parents of adolescents alter the qualities of family life for both parents and children. Rossi's study is one of the few by sociologists to incorporate social and biological influences on human development and to recognize the dialectical dynamics of simultaneous change in children and their middle-aged parents. Socialization in families becomes a two-generation process that is continual.

Whether the next decade of life-span research will continue to see the family as a useful unit of analytical distinction is not clear. Incorporation of the life-span approach into family research raises the possibility that family development may prove to be nothing more than the interactive combination of the individual developmental trajectories of coresidents. Put another way, one challenge of the new perspective for sociologists and others who have traditionally used structural features and stages of family development as analytical tools is to demonstrate that the aggregate or structural approach remains viable, given the increasing diversity (and recognition of it) of both individual patterns of development and the histories of individual families or coresidential units. This challenge is not unlike the one before students of adult development who have tended to use stage or phase models of personality change over the life course (e.g., Gould, 1978; Levinson, 1978; Vaillant, 1977). Brim and Ryff's (1980) effort to identify and classify how varieties of life events--biological, social, historical, and psychological--shape and reform the personality within both normative and nonnormative trajectories of experience may provide a necessary conceptual bridge between developmental research involving family cycles and research on the socialization of children and their parents.

The Multidisciplinary Future of Family Research

Life-span research on the historical family and its function as agents of socialization and of social change is providing new opportunities to reevaluate the modern family. In some ways there appear to be greater historical continuities, especially in economic and demographic functioning, than hitherto appreciated. In other respects, massive cohort and historical discontinuities are becoming more apparent (e.g., Brim, 1980). Life-span research on the family is inherently multidisciplinary because of its focus on individual change, social change, family process, generational relations, and bioevolutionary change. On a reduced scale of personality systems and social systems, it provides all the essential elements of the paradigm of individual and social change that underlies the entire intellectual scope of the life-span approach. As behavioral scientists carry forward the scholarly momentum of the last decade, as economists work more closely on family-related processes with historians, sociologists, and developmental psychologists, the foundation for new disciplinary breakthroughs and multilevel theory-building may be prepared. In any case, both disciplinary and multidisciplinary social science appear to be most cumulative when there is a concrete link between academic scholarship and practical problems (e.g., House, 1977). The contemporary family, in all of its myriad forms and transformations, provides that context.

PROSPECTS

Life-span research in the social and behavioral sciences is challenging old ways of thinking about the course of human development and of aging. Whether or not a new behavioral science of life-course processes emerges over the next decade in response to these developments, the promise of the next five years is for closer contact and collaboration among several existing disciplines, especially psychology, sociology, history, economics, anthropology, and biology. Within the emerging common themes and propositions of a life-span orientation, there probably will be a productive division of labor. For example, psychologists may devote themselves most intensively to uncovering ontogenetic processes and behavioral sequences that seem to have more general manifestations across historical moments. Sociologists and anthropologists may concentrate on understanding how, when, and where age becomes a basis of social organization--how events become more or less age-related or age-graded; and when and how a society becomes "age-irrelevant" (e.g., Neugarten, 1979). Economists and historians may seek an understanding of historical episodes and cohort cycles that both reflect and mold human development as a dynamic, lifelong process. Biologists may pursue cellular aging and the science of neural processes as reflections of historical changes in species longevity and cohort succession. A conscious division of labor and the recognition of common perspective should yield at least more sophisticated and ecologically valid biology, sociology, psychology, and so on.

To realize the broad academic and practical potentials of the life-span orientation over the next decade will require a new research agenda so that trustworthy generalizations can cumulate under the guidance of the new perspective. Insofar as one cohort potentially ages or develops according to its unique historical and biological circumstances, scientists must be able to compare the experiences of two or more cohorts. Replications of studies--repeating the same investigations with the same or equivalent methods--must become more common in order to monitor the course and effects of historical and institutional change on development and to assess the reciprocal influences of individual and social changes across successive cohorts. Longitudinal designs and the follow-up of panels of cohort samples throughout their lives are essential to life-span research. Intervention research and historical and cross-cultural studies must be undertaken in order to define and understand the limits and potentials of the human condition as it interacts with and transforms its context over time. This is a comprehensive research program, one involving the collaboration of several disciplines and profiting from the special skills and techniques associated with each. It calls for a sustained temporal commitment from researchers, for the organization and maintenance of longitudinal, cohort-comparative projects are both demanding and long-term. Obversely, it requires a stable base of research funding--one that recognizes both the benefits of long-run programmatic effort and the need to reassess and update the base of knowledge routinely.

It has been suggested that the social sciences were consolidated intellectually in the United States during World War II at a time when they were challenged to face the practical needs of the nation at war (e.g., House, 1977). It was a period of cross-disciplinary cooperation. The emergence of renewed interest in multidisciplinary scholarship dealing with the common themes of the life-span approach is one sign that the social sciences again may be poised to advance. Surely the practical challenges of the 1980s, which might focus this sense of new scholarly vision and common pursuit, are no less substantial than those of the 1940s.

ACKNOWLEDGMENTS

The author is grateful to the Social Science Research Council and its Committee on Life-Course Perspectives on Middle and Old Age for their collegial support and intellectual stimulation throughout the preparation of this manuscript. Particularly helpful guidance was provided by three editors, Paul B. Baltes, Orville G. Brim, Jr., and Glen H. Elder, Jr., on behalf of the committee. Lonnie Sherrod, Matilda Riley, David Kertzer, Maris Vinovskis, John Meyer, Aage Sorensen, and Brewster Smith also provided helpful insights and suggestions. Katie Knorowski and Vivien Shelanski helped to bring the text and manuscript into final form with great craft and care.

REFERENCES

Anderson, Michael
1971 Family Structure in Nineteenth Century Lancashire. Cambridge, England: Cambridge University Press.

Baltes, Margret, and E. M. Barton
1981 "Behavioral analysis of aging: a review of

the operant model and research." International Journal of Behavior Development.

Baltes, Paul B., ed.
1978 Life-Span Development and Behavior. Volume 1. New York: Academic Press.

Baltes, Paul B., and Margret M. Baltes
1980 "Plasticity and variability in psychological aging: methodological and theoretical issues." In G. Gurski, ed., Determining the Effects of Aging on the Central Nervous System. Berlin: Schering.

Baltes, Paul B., and O. G. Brim, Jr., eds.
1979 Life-Span Development and Behavior. Volume 2. New York: Academic Press.
1980 Life-Span Development and Behavior. Volume 3. New York: Academic Press.
1981 Life-Span Development and Behavior. Volume 4. New York: Academic Press.

Baltes, Paul B., and K. W. Schaie, eds.
1973 Life-Span Developmental Psychology: Personality and Socialization. New York: Academic Press.
1976 "On the plasticity of intelligence in adulthood and old age: where Horn and Donaldson fail." American Psychologist 31:720-725.

Baltes, Paul B., and Sherry L. Willis
1981 "Plasticity and enhancement of intellectual functioning in old age: Penn State's adult development and enrichment project." In F. Craik and S. Trehub, eds., Aging and Cognitive Processes. New York: Plenum Press.

Baltes, Paul B., S. W. Cornelius, A. Spiro, III, J. R. Nesselroade, and S. L. Willis
1980a "Integration vs. differentiation of fluid-crystallized intelligence in old age." Developmental Psychology 16:625-635.

Baltes, Paul B., H. Reese, and L. Lipsitt
1980b "Life-span developmental psychology." Annual Review of Psychology, 31:65-110.

Becker, Gary
1964 Human Capital: A Theoretical and Empirical Analysis. New York: Columbia University Press.
1965 "A theory of the allocation of time." Economics Journal 75:493-517.

Benedict, Ruth
1938 "Continuities and discontinuities in cultural conditioning." Psychiatry I:161-167.

Berkner, Lutz

1975 "The use and misuse of census data for the historical analysis of family structure." Journal of Interdisciplinary History 4 (Spring):721-738.

Blau, P., and O. D. Duncan

1967 The American Occupational Structure. New York: Wiley.

Block, Jack

1971 Lives Through Time. Berkeley, Calif.: Bancroft Press.

Bloom, B. S.

1964 Stability and Change in Human Characteristics. New York: Wiley.

Bowles, Samuel, and Herbert Gintis

1976 Schooling in Capitalist America. New York: Basic Books.

Brim, O. G., Jr.

1966 "Socialization through the life cycle." Pp. 1-50 in O. G. Brim, Jr., and Stanton Wheeler, eds., Socialization After Childhood. New York: Wiley.

1980 "Socialization in an unpredictable society." Plenary address to the American Sociological Association, New York.

Brim, O. G., Jr., and Jermoe Kagan, eds.

1980 Constancy and Change in Human Development. Cambridge, Mass.: Harvard University Press.

Brim, O. G., Jr., and Carol D. Ryff

1980 "On the properties of life events." Pp. 368-388 in P. B. Baltes and O. G. Brim, Jr., eds., Life-Span Development and Behavior. New York: Academic Press.

Bronfenbrenner, Urie

1979 The Ecology of Human Development. Cambridge, Mass.: Harvard University Press.

Cain, Leonard D., Jr.

1959 "The sociology of aging: a trend report and bibliography." Current Sociology 3(2):57-133.

1964 "Life course and social structure." Pp. 272-309 in R. Faris, ed., Handbook of Modern Sociology. Chicago: Rand McNally.

Clark, M., and B. G. Anderson

1967 Culture and Aging: An Anthropological Study of Older Americans. Springfield, Ill.: Charles C Thomas.

Clausen, John, ed.
1968 Socialization and Society. Boston: Little, Brown.
1972 "The life-course of individuals" Pp. 457-514 in M. W. Riley et al., eds., Aging and Society. Volume 3: A Sociology of Age Stratification. New York: Russell Sage Foundation.

Cool, Linda
1980 "Ethnicity and aging: continuity through change for elderly Corsicans." In C. Fry, ed., Aging in Culture and Society. New York: Praeger.

Cowgill, D. O., and L. D. Holmes, eds.
1972 Aging and Modernization. New York: Appleton-Century-Crofts.

Cumming, E., and W. E. Henry
1961 Growing Old: The Process of Disengagement. New York: Basic Books.

Datan, Nancy, and Leon H. Ginsberg, eds.
1975 Life-Span Developmental Psychology: Normative Life Crises. New York: Academic Press.

Datan, Nancy, and Hayne W. Reese, eds.
1977 Life-Span Developmental Psychology: Dialectical Perspectives on Experimental Research. New York: Academic Press.

Demos, John, and Sarane Boocock, eds.
1978 Turning Points: Historical and Sociological Essays on the Family. Chicago: University of Chicago Press.

Denney, N. W.
1979 "Problem solving in later adulthood: intervention research." In P. B. Baltes and O. G. Brim, Jr., eds., Life-Span Development and Behavior. Volume 2. New York: Academic Press.

DiRenzo, G. J.
1977 "Socialization, personality, and social systems." Annual Review of Sociology 3:261-295.

Dollard, John
1935 Criteria for the Life History. New Haven: Yale University Press.

Duncan, Greg. J., and James N. Morgan, eds.
1976 Five Thousand American Families: Patterns of Economic Progress. Volume VIII. Ann Arbor, Mich.: Institute for Social Research.
1980 Five Thousand American Families: Patterns of Economic Progress. Volume IV. Anne Arbor, Mich.: Institute for Social Research.

Duncan, O. D.

1966 "Path analysis: sociological examples." American Journal of Sociology 72:1-16.

1967 "Discrimination against Negroes." Annals of the American Academy of Political and Social Science 371:85-103.

Duncan, O. D., D. L. Featherman, and B. Duncan

1972 Socioeconomic Background and Advancement. New York: Seminar Press.

Dunton, Nancy, and David L. Featherman

forthcoming "Social mobility through marriage and careers: achievement over the life course." In J. Spence, ed., Achievement and Achievement Motivation: Psychological and Sociological Perspectives. San Francisco: Freeman.

Duvall, Evelyn M.

1971 Family Development. 4th edition. Philadelphia: Lippincott.

Easterlin, Richard

1980 Birth and Fortune: The Impact of Numbers on Personal Welfare. New York: Basic Books.

Eisenstadt, S.

1956 From Generation to Generation: Age Groups and Social Structure. Glencoe, Ill.: Free Press.

Elder, Glen H., Jr.

1974 Children of the Great Depression. Chicago: University of Chicago Press.

1975 "Age differentiation and the life course." Annual Review of Sociology 1:165-190.

1978 "Family history and the life course." Pp. 17-64 in T. Hareven, ed., Transitions: The Family and Life Course in Historical Perspective. New York: Academic Press.

1980 "History and the life course." In D. Bertaux, ed., Biography and Society. Beverly Hills, Calif.: Sage Publications.

forthcoming History and the Family. Unpublished paper. Department of Sociology, Cornell University.

Featherman, David L.

1980 "Schooling and occupational careers: constancy and change in worldly success." Pp. 675-738 in O. G. Brim, Jr., and Jerome Kagan, eds., Constancy and Change in Human Development. Cambridge, Mass.: Harvard University Press.

forth-coming "A history and emergent themes of the lifespan perspective in behavioral science research." In G. H. Elder, Jr., ed., Life-Course Dynamics from the 1960s to the 1980s. New York: Social Science Research Council.

Featherman, D. L., and R. M. Hauser

1978 Opportunity and Change. New York: Academic Press.

Flavell, John H.

1970 "Cognitive changes in adulthood." In L. R. Goulet and P. B. Baltes, eds., Life-Span Developmental Psychology: Research and Theory. New York: Academic Press.

Foner, Anne, and David I. Kertzer

1978 "Transitions over the life course: lessons from age-set societies." American Journal of Sociology 83(5):1081-1104.

Giddings, F. P.

1897 The Theory of Socialization. New York: Macmillan.

Glick, Paul C.

1947 "The family cycle." American Sociological Review 12:164-174.

1977 "Updating the life cycle of the family." Journal of Marriage and the Family 39:5-13.

Glick, Paul C., and Arthur J. Norton

1977 "Perspectives on the recent upturn in divorce and remarriage." Demography 10(3):301-314.

Glick, Paul C., and R. Parke, Jr.

1965 "New approaches in studying the life cycle of the family." Demography 2:187-202.

Goslin, David, ed.

1969 Handbook of Socialization Theory and Research. Chicago: Rand McNally.

Gould, Roger

1978 Transformations: Growth and Change in Adult Life. New York: Simon and Schuster.

Goulet, L. R., and P. B. Baltes, eds.

1970 Life-Span Developmental Psychology: Research and Theory. New York: Academic Press.

Hareven, Tamara

1975 "Family time and industrial time: family and work in a planned corporation town, 1900-1924." Journal of Urban History 1(3):365-385.

1977 "Family time and historical time." Daedalus 106(Spring):57-70.

1981 Industrial Time and Family Time. New York: Cambridge University Press.

Hareven, Tamara, ed.
1978 Transitions: The Family and Life Course in Historical Perspective. New York: Academic Press.

Heckman, James
1974 "Life-cycle consumption and labor supply: an explanation of the relationship between income and consumption over the life cycle." The American Economic Review 64(1):182-194.

Henry, Louis
1956 "Anciennes families genevoises." Etude demographique XVI. Paris: Processes Universitaires de France, Travaux et Documents de l'Insitut National d'Etudes Demographiques.

Hershberg, T., ed.
1981 Philadelphia: Work, Space, Family and Group Experience in the Nineteenth Century. New York: Oxford University Press.

Hill, Reuben, and Paul Mattessich
1979 "Family development theory and life-span development." Pp. 1962-204 in P. B. Baltes and O. G. Brim, Jr., eds., Life-Span Development and Behavior. Volume 2. New York: Academic Press.

Hogan, Dennis P.
1978 "The variable order of events in the life course." American Sociological Review 43(August):573-586.

1980 "The transition to adulthood as a career contingency." American Sociological Review 45(2):261-275.

Horn, J. L., and R. B. Cattell
1967 "Age differences in fluid and crystallized intelligence." Acta Psychologica 26:107-129.

Horn, J. L., and G. Donaldson
1980 "Cognitive development in adulthood." Pp. 445-529 in O. G. Brim, Jr., and J. Kagan, eds., Constancy and Change in Human Development. Cambridge, Mass.: Harvard University Press.

House, James S.
1977 "The three faces of social psychology." Sociometry 40:161-177.

Inkeles, Alex, and David Smith
1974 Becoming Modern. Cambridge, Mass.: Harvard University Press.

Jöreskog, Karl G., and Dag Sörbom
1979 Advances in Factor Analysis and Structural Equation Models. Cambridge, Mass.: Abt Books.

Kaestle, Karl, and Maris Vinovskis
1979 "From fireside to factory: school entry and school leaving in nineteenth century Massachusetts." In T. Hareven, ed., Transitions: The Family and the Life Course in Historical Perspective. New York: Academic Press.
1980 Education and Social Change in Nineteenth Century Massachusetts. New York: Cambridge University Press.

Kagan, Jerome
1980 "Perspectives on continuity." Pp. 26-74 in O. G. Brim, Jr., and J. Kagan, eds., Constancy and Change in Human Development. Cambridge, Mass.: Harvard University Press.

Kardiner, A.
1939 The Individual and His Society. New York: Columbia University Press.

Katona, George
1975 Psychological Economics. New York: Elsevier.

Katz, Michael B.
1975 The People of Hamilton, Canada West: Family and Class in a Mid-Nineteenth Century City. Cambridge, Mass.: Harvard University Press.

Keifer, C. W.
1974 "Lessons from the Issei." In J. Gubrium, ed., Late Life: Communities and Environmental Policy. Springfield, Ill.: Charles C Thomas.

Keith, Jennie, ed.
1979 The Ethnography of Old Age. Anthropological Quarterly 52(1).
1980 "The best is yet to be: towards an anthropology of age." Annual Review of Anthropology. Palo Alto, Calif.: Annual Reviews, Inc.

Kohn, Melvin
1969 Class Conformity: A Study in Values. Homewood, Ill.: Dorsey Press.

Kohn, Melvin, and Carmi Schooler
1973 "Occupational experience and psychological functioning: an assessment of reciprocal effects." American Sociological Review 38(February): 97-118.

1978 "The reciprocal effects of the substantive complexity of work and intellectual flexibility." American Journal of Sociology 84(July):24-52.

Labouvie-Vief, G.

1976 "Toward optimizing cognitive competence." Educational Gerontology 1:75-92.

1980 "Beyond formal operations: uses and limits of pure logic in life-span development." Human Development 23:141-161.

Laslett, Peter

1972 "Mean household size in England since the sixteenth century." Pp. 125-158 in P. Laslett, ed., Household and Family in Past Time. Cambridge, England: Cambridge University Press.

Lehman, N. C.

1964 "The relationship between chronological age and high level research output in physics and chemistry." Journal of Gerontology 19:157-164.

Lerner, Richard, and Nancy Busch-Rossnagel

forthcoming "Individuals as producers of their development: conceptual and empirical bases." In R. Lerner and N. Busch-Rossnagel, eds., Individuals as Producers of Their Development: A Life-Span Perspective. New York: Academic Press.

LeVine, Robert A.

1978 "Comparative notes on the life course." Pp. 287-296 in T. Hareven, ed., Transitions: The Family and the Life Course in Historical Perspective. New York: Academic Press.

Levinson, Daniel, with others

1978 The Seasons of a Man's Life. New York: Knopf.

Loomis, Charles P., and C. Horace Hamilton

1936 "Family life cycle analysis." Social Forces 15(December):225-231.

Mannheim, Karl

1952 "The problem of generations." In Essays in the Sociology of Knowledge. Translated by P. Kecskemeti. Published originally in 1928. New York: Oxford University Press.

Marshall, A.

1948 Principles of Economics. New York: Macmillan.

McCall, R. B.

1979 "The development of intellectual functioning in infancy and the prediction of later IQ." In J. Osofsky, ed., Handbook of Research in Infancy.

Modell, John, Frank Furstenberg, and Theodore Hershberg
1976 "Social change and transitions to adulthood in historical perspective." Journal of Family History 1:7-31.

Modell, John, and T. Hareven
1973 "Urbanization and the malleable household: an examination of boarding and lodging in American families." Journal of Marriage and Family 35:467-479.

Modigliani, Franco
1966 "The life cycle hypothesis of saving, the demand for wealth and the supply of capital." Social Research 33:160-217.

Myerhoff, B., and A. Simic, eds.
1978 Life's Career--Aging, Cultural Variations in Growing Old. Beverly Hills, Calif.: Sage Publications.

Nesselroade, John, and Hayne W. Reese, eds.
1973 Life-Span Developmental Psychology: Methodological Issues. New York: Academic Press.

Neugarten, Bernice
1979 Policy for the 1980's: Age or Need Entitlement? A study paper prepared for the conference, Aging: Agenda for the Eighties, sponsored by the National Journal, Washington, D.C.

Neugarten, Bernice, and Nancy Datan
1973 "Sociological perspectives on the life cycle." Pp. 53-69 in P. B. Baltes and K. W. Schaie, eds., Life-Span Developmental Psychology: Personality and Socialization. New York: Academic Press.

Neugarten, Bernice, and G. Hagestad
1976 "Age and the life course." Pp. 35-55 in R. Binstock and E. Shanas, eds., Handbook of Aging and the Social Sciences. New York: Van Nostrand.

Panel on Social Indicators
1969 Toward a Social Report. Washington, D.C.: U.S. Department of Health, Education, and Welfare.

Parsons, Talcott
1942 "Age and sex in the social structure of the United States." American Sociological Review 7(5):604-616.

Polachek, S. W.
1979 "Occupational segregation among women: theory, evidence, and a prognosis." Pp. 137-157 in C. B. Lloyd, E. S. Andrews, and C. L. Gilroy, eds., Women in the Labor Market. New York: Columbia University Press.

Pressey, Sidney, Joseph Janney, and Raymond Kuhlen
1939 Life: A Psychological Survey. New York: Harper.

Reinert, Guenther
1979 "Prolegomena to a history of life-span developmental psychology." Pp. 205-255 in P. B. Baltes and O. G. Brim, Jr., eds., Life-Span Development and Behavior. Volume 2. New York: Academic Press.

Riegel, Klaus F.
1979 Foundations of Dialectical Psychology. New York: Academic Press.

Riley, Matilda W.
1976 "Age strata in social systems." Pp. 189-217 in R. H. Binstock and E. Shanas, eds., Handbook of Aging and the Social Sciences. New York: Van Nostrand.
1981 "Age and aging: from theory generation to theory testing." Pp. 339-348 in H. B. Blalock, Jr., ed., Sociological Theory and Research: A Critical Appraisal. New York: Free Press.

Riley, Matilda W., ed.
1979 Aging from Birth to Death: Interdisciplinary Perspectives. Boulder, Colo.: Westview Press.

Riley, Matilda W., and Anne Foner
1968 Aging and Society. Volume I: An Inventory of Research Findings. New York: Russell Sage Foundation.

Riley, Matilda W., John W. Riley, Jr., and Marilyn E. Johnson, eds.
1969 Aging and Society. Volume 2: Aging and the Professions. New York: Russell Sage Foundation.

Riley, Matilda W., Marilyn Johnson, and Anne Foner, eds.
1972 Aging and Society. Volume 3: A Sociology of Age Stratification. New York: Russell Sage Foundation.

Rodin, Judith
1980 "Managing the stress of aging: the role of control and coping." Pp. 171-202 in S. Levine and H. Ursin, eds., Coping and Health. New York: Plenum Press.

Rosen, Sherwin

1977 "Human capital: a survey of empirical results." Pp. 3-38 in R. G. Ehrenberg, ed., Research in Labor Economics. Greenich, Conn.: JAI Press.

Rosow, I.

1976 "Status and role change through the life span." In R. H. Binstock and E. Shanas, eds., Handbook of Aging and the Social Sciences. New York: Van Nostrand.

Rossi, Alice S.

1980 "Aging and parenthood in the middle years." Pp. 138-207 in P. B. Baltes and O. G. Brim, Jr., eds., Life-Span Development and Behavior. Volume 3. New York: Academic Press.

Ryder, Norman B.

1965 "The cohort as a concept in the study of social change." American Sociological Review 30: 843-861.

Sapir, E.

1934 "The emergence of the concept of personality in a study of cultures." Journal of Social Psychology 5:408-415.

Schaie, K. W.

1977-1978 "Toward a stage theory of adult cognitive development." Journal of Aging and Human Development 8:129-138.

1979 "The primary mental abilities in adulthood: an exploration in the development of psychometric intelligence." In P. B. Baltes and O. G. Brim, Jr., eds., Life-Span Development and Behavior. Volume 2. New York: Academic Press.

Schaie, K. W., and Baltes, P. B.

1977 "Some faith helps to see the forest: a final comment on the Horn and Donaldson myth of the Baltes-Schaie position on adult intelligence." American Psychologist 32:1118-1120.

Sears, Robert

1980 "A new school of life span?" Contemporary Psychology 25:303-304.

Sewell, W. H., A. Haller, and A. Portes

1969 "The educational and early occupational attainment process." American Sociology Review 34:82-92.

Sewell, W. H., and R. M. Hauser

1975 Education, Occupation, and Earnings. New York: Academic Press.

1980 "The Wisconsin longitudinal study of social and psychological factors in aspirations and achievements." Research in Sociology of Education and Socialization 1:59-99.

Sewell, William H., Robert M. Hauser, and Wendy C. Wolf
1980 "Sex, schooling, and occupational status." American Journal of Sociology 86(5):551-583.

Simmons, L. W.
1945 The Role of the Aged in Primitive Society. New Haven, Conn.: Yale University Press.

Smelser, Neil J.
1959 Social Change in the Industrial Revolution. Chicago: University of Chicago Press.

Sorokin, P.
1927 Social Mobility. New York: Harper.
1941 Social and Cultural Dynamics: Basic Problems, Principles, and Methods. Volume 4. New York: American Book Company.
1947 Society, Culture, and Personality. New York: Harper and Brothers.

Spanier, Graham B., and P. C. Glick
No date "The life cycle of American families: an expanded analysis." Division of Individual and Family Studies, College of Human Development, Pennsylvania: Penn State University.

Sterns, H. L., and R. E. Sanders
1980 "Training and education of the elderly." In R. R. Turner and H. W. Reese, eds., Life-Span Developmental Psychology: Intervention. New York: Academic Press.

Stewart, F.
1977 Fundamentals of Age-Group Systems. New York: Academic Press.

Stigler, George J.
1954 "The early history of empirical studies of consumer behavior." Journal of Political Economy 62(2):95-113.

Strehler, B. L.
1977 Time, Cells, and Aging. New York: Academic Press.

Thernstrom, S.
1964 Poverty and Progress: Social Mobility in a Nineteenth Century City. Cambridge, Mass.: Harvard University Press.

Thomae, Hans
1979 "The concept of development and life-span developmental psychology." Pp. 282-312 in P.

B. Baltes and O. G. Brim, Jr., eds., Life-Span
Development and Behavior. Volume 2. New York:
Academic Press.

Thomas, W. I.
1909 Source Book for Social Origins. Boston: Badger.

Thomas, W. I., and F. Znaniecki
1918 The Polish Peasant in Europe and America. New
York: Octogon.

Uhlenberg, Peter
1979 "Demographic change and problems of the aged."
Pp. 153-166 in M. W. Riley, ed., Aging from
Birth to Death. Boulder, Colo.: Westview Press

Vaillant, George
1977 Adaptation to Life. New York: Little, Brown.

Van Dusen, R., and E. Sheldon
1976 "The changing status of American women: a life
cycle perspective." American Psychologist
31(February):106-116.

Vatuk, Sylvia
1980 "Withdrawal and disengagement as as cultural
response to aging in India." Pp. 126-148 in C.
Fry, ed., Aging in Culture and Society. New
York: Praeger.

Vinovskis, Maris
1977 "From household size to the life course."
American Behavioral Scientist 21(2):263-287.

Waring, Joan M.
1975 "Social replenishment and social change: the
problem of disordered cohort flow." American
Behavioral Scientist 19(2):237-256.

Welch, Finis
1979 Effects of Cohort Size on Earnings: The Baby
Boom Babies' Financial Bust. Unpublished manuscript.
Department of Economics, University of
California, Los Angeles.

Willis, Sherry L., and Paul B. Baltes
1980 "Intelligence in adulthood and aging:
contemporary issues." Pp. 260-272 in L. W.
Poon, ed., Aging in the 1980s: Psychological
Issues. Washington, D.C.: American Psychological
Association.

Wrigley, E. A.
1977 "Reflections on the history of the family."
Deadalus 106(Spring):71-85.

Advances in Methods for Large-Scale Surveys and Experiments

Judith M. Tanur

An army is said to march on its stomach--forward progress depending on the mundane realities of the preparation and distribution of food. Similarly, a science may be said to progress on its methods--the production of substantive knowledge, basic or applied, depending on the mundane techniques for collecting, analyzing, and interpreting data. In the centennial issue of Science, Nobel laureate Herbert A. Simon (1980:72) wrote: "An important part of the history of the social sciences over the past 100 years, and of their prospects for the future, can be written in terms of advances in the tools for empirical observations and in the growing bodies of data produced by those tools."

That is not to say that methodology alone can create or advance science--the most sophisticated method used mindlessly can produce, at best, only pyrotechnics to dazzle the uninitiated. It is the thoughtful development of specialized methods, their careful application to substantive problems, and the thorough and balanced exposition to the lay public as well as to the technically versed scientific community (stressing the limitations of the methods as well as their strengths) that shed the bright and steady light by which science can make its way forward. Our confidence in our knowledge about the social world depends to a large extent upon our confidence in the research methods employed to secure that knowledge.

This paper was commissioned by the Social Science Research Council for The National Science Foundation's Five Year Outlook on Science and Technology: 1981. Preparation of this paper was supported by NSF Contract No. PRA 8017924.

The methods used in large-scale social science research are the concern of this paper. Social science research has become a national resource. Its findings are mined to provide insights about social processes useful for both basic research and social policy applications. They inform the creation of policy. Methodological resources developed for social science research are used to design evaluations of policy and to secure information crucial to governmental decisions. In 1979 some 150 domestic assistance programs used statistical factors (mostly survey data and the census) to allocate more than $120 billion in federal funds--one-fifth of the federal budget (Wallman, 1980). The federal investment in conducting the surveys to gather the data on which these allocations were based is itself considerable, though small in relation to the amount allocated. The combined budgets of the major agencies fielding the relevant surveys were approximately three-quarters of a billion dollars in 1979.

What are these surveys that are so broadly used?

A survey is one means of gathering information about the characteristics, actions, or opinions of a large group of people, referred to as a population.[1] The population may be voters and the information sought may be their opinions of candidates or their voting intentions; the population may be recipients of food stamps and the information sought may be how the food stamps are used; the population may be consumers and the information sought may be whether they intend to purchase a new major appliance within the next year; the population may be the entire U.S. population of working age and the information sought may be the amount of unemployment. The groups interested in the results of such surveys, who therefore commission them or carry them out, vary enormously. Besides agencies of the federal government, such as the Census Bureau and the Department of Agriculture, they include commercial polling firms, organizations specializing in market research, university-based research institutes, and state and local governments.

Surveys can be classified by whether they involve a single interview or repeated interviews with the same respondents. An example of a methodologically influential one-time (or cross-sectional) survey is the Equality of Educational Opportunity Study, which collected data on some 570,000 school pupils, 60,000 teachers, and 4,000 schools. The study asked questions about the effect of school facilities, teacher characteristics, and home situations on students' educational performance. Its analyses and reanalyses, discussion and controversy generated much heat but also shed much methodological light on succeeding studies (see Mosteller and Moynihan, 1972). Surveys that interview the same respondents repeatedly are called longitudinal or panel surveys. They have major advantages over one-time surveys in their ability to follow individual changes over time and thus illuminate the social processes that are at work. Thus the Panel Study of Income Dynamics (Morgan, 1977) can describe those Americans who are persistently poor (that is, below the poverty line year after year) and contrast them with families who dip below the poverty line only once in a decade of interviewing. Similarly, the Parnes Study, also called the National Longitudinal Surveys (Bielby et al., 1977), can investigate the long-term effects of chronic teenage unemployment on future labor force participation.

---

[1] The American Statistical Association has recently published a booklet entitled What is a Survey? that clearly explicates the variety, purposes, and methods of basic survey research (see Ferber et al., 1980).

Surveys can also be classified as to whether they are seeking information about a system as it stands, perhaps to establish a baseline against which to measure the effect of a policy change, or about the impact of a program change after it has been implemented, either in full or on an experimental basis. The Current Population Survey, carried out each month by the Census Bureau to measure (among other things) the extent of employment and unemployment, is an example of a survey that measures a current condition, as was the survey conducted to measure equality of educational opportunity. Surveys designed to measure the impact of a program or policy change can be roughly divided into two groups: those embedded in quasi-experiments and those embedded in true social experiments.

Quasi-experiments[2] using surveys fall into two broad classes. One kind is the interrupted time-series design, in which data are collected from the same people or from the same population for a considerable period of time before and after some policy change or intervention occurs. The data from the "before" period give some indication of the trends in the phenomenon and its variability in the absence of any intervention. These provide a yardstick with which to measure the effect of the intervention and to gauge whether a discontinuity in the time series at the point of intervention represents a real change. This use of surveys has had little large-scale application to date because no sufficiently long, valid, and credible time series have existed. The massive data collection and documentation efforts detailed in this paper may well yield such time series for future use.

---

[2] Useful catalogues of designs for quasi-experiments and the particular threats to validity they defend against have appeared widely in the literature; see, for example, Campbell and Stanley (1963); Campbell (1978); and Cook and Campbell (1979).

The second kind of quasi-experimental use of surveys seeks data from people in a new program as well as from control groups. The control groups are designed to be as similar to the program groups as possible on the variables that are expected to influence responses to programs or to be important as alternative explanations for changes in the program groups. An example is the study of the impact of Project Head Start, conducted by the Westinghouse Learning Corporation and Ohio University under contract with the Office of Economic Opportunity in 1968-1969. The study matched 1,980 first-, second-, and third-grade students from 104 Head Start centers with a control group of nonparticipating children from the same schools in evaluating the program (Granger et al., 1969). The two groups were matched on the key variables of age, sex, race or ethnicity, and kindergarten attendance. The socioeconomic status of the two groups was made as comparable as possible by the use of a statistical technique called analysis of covariance. Small but statistically significant differences on a few of the outcome measures in favor of the full-year Head Start group were found.

Quasi-experiments are fraught with inferential dangers. A well-designed and analyzed quasi-experiment makes every effort to separate the effects due to the program from effects due to other causes. For example, such investigations must always ask whether self-selection into the program by those most likely to benefit from it causes the program to appear more effective than it would be if offered to the general population. Conversely, selection of those most in need of the program may also select those least able to benefit and thus make the program appear less effective than it should. However much care is exercised, such isolation of program effects is never completely possible in quasi-experiments. Thus, unambiguous causal statements--bold statements that the program caused the outcome--cannot be made from quasi-experiments.

It is for this reason that true social experiments, the other application of surveys to evaluating the impact of program change or innovation, are mounted. In these a proposed policy innovation is implemented experimentally among a randomly chosen group of potential participants, with another randomly chosen group serving as a control to evaluate outcomes in the absence of the program. It is the act of randomization that makes the control and the experimental groups the same before the program is introduced and thus permits probabilistic assessment of whether any subsequent differences between them are effects of the program.

The earliest of these large-scale true social experiments was the New Jersey Negative Income Tax Experiment, and it was followed by several other income maintenance experiments across the country. The purpose of these studies was to find out whether a government-supplied income supplement to poor people would reduce their incentive to work. The findings indicate that little disincentive is created for primary wage earners, but slightly more exists for secondary wage earners. Other large-scale social experiments have investigated the effects of offering housing allowances and differing health insurance plans.

Surveys in their various forms are one of the tools that Herbert Simon refers to in the passage quoted earlier in this essay as to providing a growing body of data. But Simon goes on to say (1980:72): "It is perhaps not important that we have more information than our ancestors; it is vitally important that we have better information. A major part of the effort of trained social scientists has gone into improving our techniques for making the kinds of measurements that I have just enumerated [essentially survey data]."

Sampling and standardization are two key aspects of proper surveys for which methodological advances make possible the gathering of better information. A survey is used to obtain information, not from every member of the population, but from a sample selected by the use of probability methods. If the sample is drawn in a properly random manner (not haphazardly, for example, or by the use of volunteers or those conveniently available), then the results of the survey can be generalized to the population from which the sample was drawn. The second key attribute of a proper survey is that its procedures are standardized--it uses prescribed forms of questions and standardized methods of asking them.

In one way or another, sampling and standardization to obtain valid measurement are the themes of this paper. What are the methods that social scientists and statisticians have devised for making surveys yield better information? Probability sampling methods have a long history of theoretical development, but methodological attention has turned recently to those factors that can destroy the value of information from surveys even when probability samples are employed--problems of nonresponse and mistaken responses. Peeling away the variability originating from these extraneous sources purifies our information, giving us more confidence in its validity.

What are the effects of different decisions about the standardized procedures for a particular survey? Does it matter whether the interviewing is done in person or on the telephone? Does the form, context, or ordering of questions make a difference in the estimates prepared from the surveys? As research is done to answer these and related questions, we learn more about the validity and generalizability of the information supplied by surveys and are more able to improve them.

This paper discusses some of the methods that have been and are being developed to reduce the fuzziness of the knowledge gained from large-scale surveys and experiments. Ideally we should like to have a coherent and comprehensive theory of human behavior and a foolproof machine for measuring the effects brought about by our precisely specified causes. But usually our theories are stated only in broad and general terms. We often have only approximate ideas about causality. Our ability to measure effects is limited by the extraneous variability in measurements brought about by the process of sampling as well as by the standardization decisions made in any particular case. Our ability to measure effects is also limited by people's insistence on acting like human beings--refusing sometimes to answer our questions, insisting sometimes on their own interpretations of meanings rather than the ones we have in mind, and so on. It is to the separation of these extraneous sources of variability from the actual measurement of the phenomena of interest that we turn as we examine the concept of total survey variability. In the process we hope to see where new advances will arise. It is difficult (perhaps impossible) to shine a beacon into the future--but if we illuminate the recent past we may perceive the methodological advances that project their influential images onto the years ahead.

TOTAL SURVEY VARIABILITY

Although surveys and experiments are conducted using samples of individuals, their purpose is to learn more about the broader population from which the sample is taken. That information may range from the answers to relatively simple questions, such as "What proportion of the population is female?", through more difficult or sensitive ones, such as "What is the average annual expenditure for medical care for members of the population?", to conceptually complicated ones, such as "What effect on the incentive to work among the members of this population would an income supplement have?" In all cases, what is of interest is correct answers to these questions for the population sampled, not answers that are correct only for the people surveyed, nor answers that are incorrect even for the people surveyed and hence, of course, incorrect for the population as well.

As the results of surveys become more important for society--for example, as unemployment statistics from the Current Population Survey become the basis for distributing federal funds, as citizens' expressed opinions on issues of the day come to shape the platforms (and later the policies) of political candidates, as poll results, rightly or wrongly, become the basis for including or excluding a third-party presidential candidate from nationally televised campaign debates--it becomes more and more important to make these results as accurate as possible. Efforts to improve the accuracy of surveys (and other data collection methods) focus on the sources of inaccuracies; the underlying assumption is that if such sources can be identified they can eventually be controlled or, at the very least, their effects can be taken into account in the interpretation of results.

To address sources of inaccuracy many investigators use the concept of total survey error, and several models have been developed to operationalize the concept (e.g., Lessler, in press; see Mosteller, 1978, for a simple technical exposition of the Census Bureau Model). I shall use the blueprints of these models to guide this exploration of the effects of variability in surveys.

This is the first of many times in this paper that the notion of a model appears. What is a model? It is a formal expression of a theory or a set of causes that the proposer regards as having generated the observed data. In statistics such a model is usually expressed in symbols--and thus is a mathematical model. While architects and engineers construct scale models of their projects, the model of the statistician is not this sort of scaled-down but concrete representation of an object. It is rather a model of an abstract process, usually greatly simplified, and used frequently to explore how varying the inputs affects the outputs of that process. A better analogy than the model of the architect or the engineer is the animal model of the biologist. In a toxological study using an animal model, the assumption is made that the vital processes of the animal are sufficiently like those of a human being that information about a drug's effects on the animal will have some value in understanding its effects on humans. There is, however, no assumption that the animal is "just like" a human being. Similarly, a statistical model is not "just like" the process it represents; rather it abstracts the most salient elements of the process for study.

Statistical models serve as the core of surveys and experiments. The income maintenance experiments were guided by a model that suggests that earned income depends on the amount of the income supplement and other variables. The purpose of the experiments was to find out how this dependence is expressed. Certain quantities in the model (e.g., amount of income supplement given) were known for each participant; other quantities in the model--the parameters--were unknown, and the purpose of the experiment was to estimate them. Does doubling an income supplement reduce earned income by one-third? By one-half? More? Less?

It is largely through the proposing, estimating, testing, refining, reestimating, retesting, and rerefining of models that social science, its methods, and its applications advance. For example, the massive systems of structural equations that model economic processes are the backbone of econometrics; less ambitious structural equation models are used to model smaller social processes. We shall encounter models of labor force participation, marital dissolution, and other processes in this paper, but for now let us return to a specific kind of model, that of total survey error.

Total survey error models partition the total variation in survey responses into components that can be studied separately. These components include sampling variability, response effects, nonresponse effects, and their combinations. The goal is to measure and control the total error by providing a mathematical framework for examining separate sources of error. When a real survey is evaluated using the concept of total survey error, the effect of each component is gauged, and then the separate effects are synthesized to arrive at a statement about the accuracy of the entire survey. Similarly, this paper examines in turn the major components of total survey variability--sampling variability, response effects, and nonresponse effects--and then looks at some progress being made in synthesizing these ideas in measuring the accuracy of surveys.

Sampling Variability

For concreteness in defining sampling variability or sampling error, let us think about estimating the average income for the population of a city from a fully realized probability sample. (In a probability sample everyone on a list called the "frame," which defines the population, has a known nonzero probability of being included in the sample, and all samples have known probabilities of being chosen. For our example, the frame might be a list of all residents of the city, and all residents might be assigned equal possibilities of inclusion in the sample. The probability sample is fully realized--in this case--if all people chosen for the sample respond with data on the item asking for income.) Different samples would, of course, include different people and thus would be likely to yield slightly different results when average income is calculated. Conceptually, the measurement of this variation over samples is the measurement of sampling error. For a particular sample, sampling error is defined as the difference between the estimate of the average income of the population derived from the sample and the true average income of the population that would have been obtained if everyone listed in the frame had been asked the same question about income at the same time that the people in the sample were asked it. Clearly the size of the error in any given sample is unknown--for if the true population average income were available for comparison, there would be little point in carrying out the survey in the first place. Nevertheless, the variability associated with these errors over all possible samples is known from statistical theory, and it can be estimated from the variation in a particular set of sample data and the size of the sample. In particular, as the number of people in the sample (n) increases, the probability that many commonly used sample statistics (sample average income in our example) will be near to the population parameters they are designed to estimate (population average income, in our example) becomes larger. In fact, in simple random sampling, the standard error, a customary measure of sampling error, decreases proportionally to $1/\sqrt{n}$.

Thus, suppose the Current Population Survey questioned a simple random sample of 400 people (it actually uses much more elaborate sample designs and questions many more people). If the unemployment rate was found to be 10 percent, then the estimated standard error would be about 2 percentage points. Furthermore, in 95 of 100 cases the results based on the sample would differ no more than 4 percentage points in either direction from what would have been found by interviewing all eligible adults. If, however, the survey questioned 40,000 people and still found the unemployment rate to be 10 percent, then the estimated standard error would be reduced by a factor of 10 as would the length of the "95 percent confidence interval" described above.

Nonsampling Variability

But all this assumes an ideal world--among other things, it assumes that the frame is an accurate representation of the population to which we want to generalize, that everyone chosen for the sample provides data on income, that the researcher and the respondent share the same definition of "income," that respondents remember correctly and tell the truth, and that nobody makes a mistake in copying down the answer. It is to these nonsampling errors that much interest has recently turned, as results of surveys are taken seriously by the public and policy makers, for in some ways they are harder to understand and control than sampling errors. They cannot, for instance, be decreased just by increasing the size of the sample; as James A. Davis (1975:42) has put it, "$\sqrt{n}$ wrongs do not make a right."

Nonsampling variability or errors can be subdivided into nonresponse variability or errors (people are left out of the frame, left out of the sample, or do not answer specific questions) and response or "measurement" variability or effects or errors (answers are obtained, but are in some sense "wrong"). We will first consider response errors, a problem that has attracted a good deal of attention in recent years.[3]

Response Effects

Different types of questions are asked in surveys. There are factual or behavioral questions ("How old are you?" "Have you ever been arrested?") for which there is a "true" answer that can, at least in theory, be ascertained for checking the survey response against. At the other end of a scale of concreteness are attitude questions ("Do you feel that the President's policies are sound?" "Would you install insulation in your home if fuel oil prices tripled?") for which there is no external source of a "true" answer. There are continuing debates about whether the concept of "true" answers is even applicable in such cases, and an extensive literature on the match--or lack of match--between expressed attitudes and actual behaviors (see, e.g., Deutscher, 1973). In addition, there are questions that are indeed behavioral ("Have you been the victim of an unreported crime this month?") for which no easy outside verification is possible. For current purposes a distinction between factual and attitudinal questions is helpful. With factual questions we may certainly speak of response "errors" when the answer in the survey does not match a publicly recorded fact; with attitude questions we should speak of response "effects" if two different methods in a survey produce two different answers. That is, if a higher percentage of respondents answer "yes" to the question "Do you agree with the President's policies?" than answer "no" to the reversed question "Do you disagree with the President's policies?" we have a response effect attributable to question wording.

---

[3] A glossary of terms used in the discussion of nonsampling errors has been prepared, as have several excellent reviews of the literature and bibliographies in recent years (for example, B. A. Bailar, 1976; Bradburn, 1978; Dalenius, 1977; Deighton et al., 1978; Kahn and Cannell, 1978; Mosteller, 1978; Sudman and Bradburn, 1974), often in connection with continuing programs on research on survey methodology. A special issue of Sociological Methods and Research on survey design and analysis, concentrating on errors in surveys, appeared in November 1977. Two panels sponsored by the Committee on National Statistics of the National Research Council are completing work in the general area.

Three broad classes of response effects can be identified (Sudman and Bradburn, 1974): (1) those originating with characteristics of the respondent, (2) those originating with characteristics of the interviewer (or with the interaction between characteristics of the interviewer and those of the respondent), and (3) those originating in the social situation of the interview. This threefold division is followed here, although the categories and the variables within them interact. For example, a question form that gives valid data in a face-to-face situation may be inappropriate in a mail survey.

Respondent Effects Differences in respondent characteristics in general ought to create real response differences, not ones that might be called "errors." Thus the whole point of a survey, for example, might be to find out if respondents who differ on whether they live with a spouse or live separately also differ in income. Respondents may also possess other characteristics that predispose them to give particular sorts of responses, such as a need for approval, a propensity to acquiesce, or a wish to give socially desirable answers. These predispositions, unrelated to the content of the researcher's question, are called "response sets." Thus if unmarried heads of households tend to give more socially desirable responses than do married ones, they may exaggerate their income, and the true relationships between marital status and income would be obscured. Measures of such a "response set" are hence often included in questionnaires so that their impact can be controlled. But some recent research indicates (Bradburn et al., 1979) that "response sets" may not be artifacts to be eliminated but real personality traits. People who score high on these measures seem to live in limited social environments. They report low levels of behaviors such as sociability, drinking, intoxication, and marijuana use, not because they "are manipulating the image they present in the interview situation" but because they "have different life experiences and behave differently from persons with lower scores" (p. 103).[4]

---

[4] The Panel on Survey Measurement of Subjective Phenomena of the Committee on National Statistics carried out a

Memory is another respondent variable. In factual questions, a respondent must be able to remember correctly in order to give an accurate answer. Two kinds of memory errors can be distinguished: forgetting and what has come to be known as the telescoping of time. In the latter, events, purchases, victimizations, etc., are reported as happening more recently than they actually did. (This moving of events to more recent times is the usual meaning of telescoping; there is some evidence, however, that telescoping may sometimes move events into the more distant past.)

These phenomena work in opposite directions in producing response errors; forgetting leads to underreporting the number of events in a time period, and telescoping typically leads to overreporting. Forgetting can be minimized by using such memory jogs as "aided recall" (perhaps better called recognition) in which the respondent is read or shown a list of the events that may have happened and asked to indicate with a yes or no answer whether indeed they have, but this may increase telescoping. The encouragement of respondents to take the time to find records of expenditures on such items as health care and home improvements offers increased accuracy and controls telescoping, but it is of little use when records are fragmentary or nonexistent.

To control telescoping a technique called "bounded recall" has been useful in panel studies in which respondents are interviewed repeatedly (Neter and Waksberg, 1964). At the start of the second and subsequent interviews, respondents are reminded of what they have previously reported and asked what has occurred since those events. Clearly the interviewer needs an easily available and extensive fund of information on the respondent for this technique to be used extensively, and in this connection Computer Assisted Telephone Interviewing (see below) offers tremendous potential benefits. As the length of time between interviews increases, forgetting increases but telescoping decreases; conversely, as the amount of time between interviews decreases, telescoping increases and forgetting decreases. This relationship suggests that there might be an optimal spacing between interviews, so that the effects of the two phenomena tend to cancel out (see Sudman and Bradburn, 1973).

---

further review of the literature on response sets (Turner and Martin, forthcoming).

Problems with faulty memory can be avoided by asking people to keep diaries of their time use. This approach has been employed in basic research on working both in the home and outside it (e.g., Berk and Berk, 1979; and the Time-Use Survey being conducted at the University of Michigan: see, for example, M. Hill and F. T. Juster, 1979; Stafford and Duncan, 1979). Surveys by the Census Bureau have used expenditure diaries to investigate purchase of small, easily forgotten items, and these data are used by the Bureau of Labor Statistics to help decide when items included in the Consumer Price Index ought to be revised (Hoff and Thompson, 1980). Diaries of gasoline purchases have been used by the Energy Information Agency to supplement the data gathered from the residential energy consumption survey (Thompson et al., 1980). But diaries are costly, possibly incomplete, and respondent cooperation is difficult to obtain and often deteriorates with time (Kalton and Schuman, 1980).

Some of these problems can at least be addressed. For example, incentive payments have increased the completion rate of diaries (Thompson et al., 1980), and tape recording can be effective for groups who may have difficulty writing diaries (Sudman and Ferber, 1971). Another approach is to employ electronic "beepers." One group of researchers gave a sample of adolescents these electronic paging devices, through which signals were transmitted at random times (Csikszentmihalyi et al., 1977). The youths were quite cooperative in pausing in their activities to fill out a brief questionnaire about what they were doing, with whom, and how they felt about it. (Most time was passed watching TV or in conversation with peers; only 18 percent of their time was spent studying or working.) This technique seems to have wider applicability. Also used in the Michigan Time-Use Survey, it was found to give results comparable to those obtained by more usual diary methods.

Interviewer Effects The second sort of response effects are those due to interviewer characteristics or to the interaction of those characteristics with those of the respondents. The change in the U.S. census after 1950 primarily to self-reporting came about because analysis showed that enumerator effects, while not themselves terribly large, constituted a major part of the total variability of the census.

A recent review of the literature (Sudman and Bradburn, 1974), however, found evidence of only weak effects in this category. Matches between interviewer and respondent on such characteristics as gender or race tend to affect only those questions that relate directly to the matched variable. Thus blacks tend to give more militant answers to black interviewers than to white ones--raising the question of which answer is closer to the "true" attitude or behavior. When such an interaction is thought to be important, the sample can be split between matched and unmatched interviewer-respondent pairs and any response effects that arise reported as part of the data.

Interview Effects Far more important than the previous two categories in creating response effects are variables having to do with the task confronting the respondent and interviewer and the social situation in which they find themselves.

Comprehension and communication are the first interview variables: the investigator and the respondent must understand the question and the possible answers in the same way. Some startling examples of misunderstanding have been reported (Kalton and Schuman, 1980). Respondents ignored a carefully worded definition of "a room" when reporting on the number of rooms in their home (after all, they knew what a room is, and nobody had to tell them how to count). And only 1 of 246 respondents to the question "What proportion of your evening viewing time do you spend watching news programs?" could specify how to work out the proportion. (Perhaps someone does in fact have to tell respondents how to do complicated counting.) Fitting question wording to respondents' understanding, requesting clarification, and asking parallel questions with consistency checks move in the direction of improving comprehension and communication.

Mode of presentation is a second interview variable. Although the popular image of a survey taker is probably that of an earnest female interviewer ringing the doorbell of one of the chosen, in many surveys no interviewer appears at all. Some are conducted by mail, with the respondent filling out the questionnaire unassisted, and many are conducted by telephone. Mail and telephone surveys are less expensive than those conducted in person, so it becomes important to find out whether they produce differential response effects. No method has been shown to give clearly superior results for all kinds of questions (Sudman and Bradburn, 1974).

There are essentially no differences between telephone and in-person modes for nonsensitive questions (Groves and Kahn, 1979), nor even for somewhat sensitive ones for which external validity checks are possible. For example, while 57.1 percent of the noninstitutionalized U.S. population actually voted in the 1972 presidential election, overreporting of voting occurred at almost the same level among those interviewed in person (66.6 percent) and those interviewed by phone (69.1 percent) in the Groves and Kahn study. There is some evidence of greater validity in self-reporting mail forms on sensitive questions about such matters as minor lawbreaking.

Telephone interviews have to give up the visual aids often used in face-to-face interviews, for example, when the respondents are handed a card and asked to choose a response category. This procedure allows respondents to say a letter rather than directly state an income in dollars to the interviewer. But some researchers (Durako and McKenna, 1980) have found it possible to mail out visual aids in advance of an appointment for a phone interview. Only small differences in the distributions of answers from the two modes, with and without visual aids, however, have been found.

Open-ended questions (which the respondents must answer in their own words) are answered differently on the phone than in person; on the phone answers tend to be shorter and there tend to be fewer multiple answers. In an experiment done in connection with the National Crime Survey, respondents who were interviewed mainly by telephone reported themselves victims of fewer small thefts than those who were interviewed mainly in person. The effect was strongest for males and for those between ages 25 and 49. Thus an increase in telephone interviewing would change comparisons between population subgroups in the National Crime Survey (Woltman et al., 1980).

Idiosyncrasies of particular interviewers tend to have more effect in phone surveys, because each interviewer does more interviews (Groves and Magilavy, 1980). Mail and telephone interviews also sacrifice traditional interviewer skills: recognizing puzzlement from nonverbal cues and giving reassuring nonverbal messages in return; being able to code the ethnicity and social class of respondents; and being able to report on distracting influences present at the interview that may have response effects (for example, victimization by a member of one's family is unlikely to be reported while that family member is present). Telephone interviewing, at least in single-stage procedures, also sacrifices the ability to match the gender and/or race of the interviewer with those of the respondent, but as we have seen, lack of such matching produces response effects only on the questions to which such attributes are most salient.

One mode of presentation, which tends to increase anonymity because it never forces respondents to tell whether a sensitive question has been answered, ought to decrease response effects. The randomized response technique (Warner, 1965) requires a respondent to do some kind of randomization (e.g., toss a coin) to determine whether the sensitive question or an innocuous one is to be answered. (The respondent simply answers the question, but does not reveal which question he or she is answering.) Simple probability calculations then give an estimate of the number in the sample who agreed with the sensitive question without revealing which respondents did so. The technique has been found to reduce distorted responses to socially undesirable questions (that one would expect to be underreported), but to be ineffective in reducing distortion to questions dealing with socially desirable behavior (that one would expect to be overreported) (Locander et al., 1974).[5]

Still another mode of presentation designed in part to increase anonymity and hence increase response accuracy is called network sampling. Individuals, rather than being asked about their own behavior or characteristics, are asked about behaviors or characteristics of their friends or relatives (Sirken, 1975; Sudman et al., 1977).

In summary, the usual modes of presentation introduce few response effects on nonsensitive questions; with more sensitive questions, however, the more anonymous modes seem to elicit more valid responses. In addition, the shortness of telephone interviews may permit respondents to decide that some incidents are too trivial to mention. The Current Population Survey uses both telephone and in-person interviews, and there is little or no evidence that these different modes create response effects on statistics about employment and unemployment, although further research on the topic has been called for (Brooks and Bailar, 1978). The 1980 census experimented with telephone rather than in-person follow-up for a sample of those who did not mail back the census forms in order to compare the modes of completeness of data, costs, and interviewer attrition (B. A. Bailar and S. Miskura, 1980); the results of this trial are not yet in.

---

[5] Some have argued that the technique is confusing to both interviewers and respondents, and it is certainly true that it reduces the sample size by soliciting answers from only a fraction of the respondents. While the evidence on its efficacy is not conclusive, in 16 studies that compared randomized response with some standard, 9 showed a notable reduction in response error (Boruch and Cecil, 1979). An important issue is that when the percentage of people engaging in the sensitive behavior is small, the technique seems sensitive to reporting errors in the innocuous question (Shimizu and Bonham, 1978).

Even while investigators are attempting to understand the response effects connected with traditional modes of interviewing, within the last decade a new mode has been developed, and the response effects that it may introduce must take their place on the research agenda. This new mode, which may turn out to be a major innovation in interviewing, is Computer Assisted Telephone Interviewing (CATI).

Rather than reading from a printed questionnaire, the interviewer reads questions from the screen of a cathode ray tube attached to a computer terminal and records answers by typing them in on the keyboard of the terminal. Because a computer is involved, CATI offers greatly increased flexibility from beginning to end of the interviewing process. Interviewers can be presented with sample telephone numbers to be called in random order, callbacks can be automatically scheduled, and respondent selection probabilities can be altered as interviewing progresses (Dutka and Frankel, 1980; Roshwalb, Spector, and Madansky, 1979).

In a printed questionnaire instructions to the interviewer about which questions to ask of which respondent can get very complicated very quickly, and it has been common practice to allow no more than four levels of contingency (e.g., ask this question only of males, over 28, with children, and no military service). Using CATI, because the computer is programmed to do this "branching," as many as 17 levels of contingency have been used (e.g., the California peak load pricing experiment; Lebby, 1980). Contingent questioning can be used to explore successively more sensitive areas, thus providing more information and less nonresponse; respondents typically drop off only after supplying at least some information. Information from earlier in the interview can be introduced into subsequent questions, as can material from earlier interviews with the same respondent if the study is longitudinal. Question wording can be tailored to the respondent, for example, to the appropriate level of education, thus bringing the meaning of the question as intended by the researcher and as understood by the respondent into closer correspondence than is usually possible with a structured questionnaire.

Most systems for CATI can also do calculations to provide sample statistics as data arrive, and sample sizes can be determined sequentially. Errors are reduced, because the operations of data coding and entry are short-circuited and because most systems are programmed to recognize wild or inconsistent values and request correction on the spot.

CATI may well offer the opportunity to test much of the conventional wisdom of professional survey researchers (Freeman, 1980). For example, because question order can be easily, independently, and automatically randomized, and records can be kept automatically of which respondents receive what order, experiments on question ordering can be carried out routinely, as can experiments on the effect of the order in which the interviewer reads the possible responses to questions.

Switching from hard copy questionnaires to CATI creates some problems: Flexibility that is needed but not anticipated by the system designer is difficult to achieve; interviewer training differs from what is traditionally done; and currently systems from different installations are incompatible (Groves et al., 1980; Shanks, 1980). It is not clear at this time whether these are the early growing pains of a new technology or more permanent faults.

There is speculation that CATI, if used imaginatively, can represent a quantum leap in technology. For example, there has always been a tension in the construction of survey instruments between the canons of good measurement that dictate multiple indicators--as in a battery of questions measuring a psychological trait--and constraints of time and respondent patience that dictate the use of single questions or at most a few indicators. One could conceive of asking a question or two to determine the approximate scale location of a respondent (e.g., toward the conservative end of the scale), then using the flexibility of CATI to choose further questions tailored to particular respondents, placing them at more precise scale locations. Screening in telephone interviews for routine demographic and other characteristics will, with some regularity, turn up respondents who are of special interest for policy or other reasons, e.g., members of sparse groups (young Chicanos, sufferers from a rare disease). If interviews are ongoing using CATI for several studies, it would be possible to program the system to introduce a module of questions pertinent to the research concerns about the sparse group into an ongoing interview whenever a member of that group is found, thus gradually gathering a sample of sufficient size for generalizing. (But see the discussion below of the response effects due to questionnaire context; such problems may make data gathered in this way less attractive than they seem at first glance.) Finally, the notion of compressing CATI into a microprocessor so as to make the "questionnaire" portable and playable through a TV screen in a respondent's own home has been suggested (Lebby, 1980; Shanks, personal communication). Such a procedure might capitalize on respondents' positive reaction on being informed that a computer is involved in the interaction; it might also be a device for assuring respondents of confidentiality of survey data, for respondents could interact with the CATI system without the intervention of the interviewer, presumably secure in the knowledge that no one would see their identified data.

Question form--the art of question writing and questionnaire construction--has been described for years in texts, manuals, and word-of-mouth instruction. The scientific study of response effects produced by these variables also has a long history, but the more recent availability of survey archives and the increasing seriousness with which survey results are regarded have inspired a new flowering of research. (Sudman and Bradburn, 1974, provide a detailed review, as do Kalton and Schuman, 1980.)

Open-ended questions (e.g., "What is your opinion of the President's handling of the crisis in Iran?") have long been believed to give more accurate information on respondents' attitudes than closed-ended ones ("Do you think the President's handling of the crisis in Iran is excellent, fair, poor, or terrible?"). Current opinion is that neither form has a clear superiority overall. Open-ended questions are clearly needed, however, in at least two situations: first, when the salience of an issue to the respondent is being investigated, so that the respondent's words indicate the thought invested in the topic; and second, in the early stages of questionnaire construction, when the freely chosen wording of pretest respondents is crucial to the construction of response categories to be used in the closed-ended questions for the bulk of the survey.

Long questions are in bad repute for slowing down the pace of the interview and supposedly confusing respondents. Recent studies, however, have experimented with lengthening questions by adding redundant or irrelevant material without complicating them. (For example, instead of "What health problems have you had in the past year?" one might say "The next question asks about health problems during the last year. This is something we ask everyone in the survey. What health problems have you had in the past year?") The result is sometimes a longer answer from the respondent and frequently a more accurate one, in the sense that more events are reported. Longer answers seem to be given even for shorter questions when they are mixed in with long ones in a questionnaire (Cannell, 1977). Perhaps the interviewer is both modeling and reinforcing longer answers by asking longer questions.

It has long been believed that although changing the form of the question may change the distribution of respondents among the response categories (e.g., if one asks "Are you in favor of ERA?" instead of "What is your opinion of ERA, are you in favor, neutral, or opposed?", one is likely to get different percentages of respondents reporting themselves in favor of the amendment), the correlation between answers to such a question and other variables would not change with the form of question. This is the notion of "form-independent correlation." Recently, as part of a continuing program of research on question effects, questions on attitudes about foreign governments sometimes included an option of "no opinion" and sometimes required respondents to volunteer that they had no opinion if that was the case. Not only did the percentage of respondents reporting "no opinion" increase when the alternative was explicitly offered, as expected, but also the correlation between items asking opinions of different foreign governments changed (Schuman and Presser, 1978). Similarly, the correlation between changes in interest in religion and changes in attendance at religious services appeared stronger when the two questions had similar response categories (Duncan and Schuman, 1980).

The context in which a question is asked--the ordering of questions, the inclusion of other questions, the very arrangement of a questionnaire--can produce response effects. The ordering of the questions within a questionnaire may produce effects through several mechanisms (Sudman and Bradburn, 1974): (1) Order may influence the salience of topics (with topics of low salience being most affected because it is easier to increase salience than to reduce it). (2) If there is overlap between questions, respondents may be reluctant to be redundant and repeat details they have given earlier. (3) An urge to consistency might cause answers to earlier questions to influence later ones--respondents express less confidence in institutions when such questions are asked after questions about political alienation than when they are asked before (Turner and Krauss, 1978). (4) Later questions in a lengthy questionnaire may be answered in a perfunctory manner because of fatigue. In a variant of this problem, fewer incidents of victimization were reported if the questionnaire was structured so that detailed information for each incident was requested immediately after the incident was mentioned than if the respondent was encouraged to list all incidents of victimization before being asked to describe any one in detail (Biderman et al., 1967). (5) The opposite of a fatigue effect may occur, so that the rapport between respondent and interviewer may grow as the interview proceeds--thus, sensitive or threatening questions are often placed late in an interview when rapport is presumably high.

In particular, questionnaire context may well affect responses to questions that have few everyday implications (e.g., "What is your opinion of U.S. foreign policy?" versus "How many children do you plan to have?"), responses to questions with ambiguous response categories (e.g., very happy, pretty happy, versus one child, two children, etc.), and responses to questions on somewhat vague or amorphous concepts (e.g., attempted assault versus actual assault as forms of victimization; Turner, 1981). Several experiments, across survey organizations but at approximately the same time, are currently addressing these contextual effects. (The project is being encouraged by the Panel on Survey Measurement of Subjective Phenomena of the Committee on National Statistics of the National Research Council.)

In addition, light could perhaps be shed on the problem of contextual dependence if investigators were to switch focus from the question, asking which forms are susceptible to contextual effects, to the individual, or perhaps to the type of individual, asking what sort of person is affected by context. Is it the better or more poorly educated person whose thinking changes with the context in which a question is asked? Is it those who have given the matter a great deal of thought or those who have not yet thought deeply about the issues? Of course, these variables are more difficult to study than those relating to types of questions. Variables relating to individuals (other than demographic variables) are most logically studied using a test-retest design, which is difficult to administer and has artifactual problems of its own, rather than using the split-ballot design typically employed to investigate the effects of context variations over aggregates of people.

The very appearance of a self-report questionnaire may produce response effects, especially inaccuracy. The 1980 census experimental program sent out variants of the usual census form that were "people-oriented" in contrast to the standard form, which is "computer-oriented" (B. A. Bailar and S. Miskura, 1980). These forms, in view of their additional data transcription costs and the risks of error they entail, will have to show major advantages over the machine-readable questionnaire in mailback rate and data completeness if their use is to be justified.

Current Research Prospects The comparison between modes of interviewing is an area in which we can expect more research and perhaps more definitive results over the next few years. In particular, the advantages and drawbacks of CATI will be explored. Systems are currently being used or developed in surveys by commercial firms, university-based research centers, and the U.S. Bureau of the Census (Nicholls et al., 1980).[6] The branching flexibility of CATI will produce data that are themselves hierarchical. Statistical methods designed to deal with such data sets do not yet exist; we would expect that the existence of the data sets would stimulate development of the methodology.

The work of the Panel on Survey Measurement of Subjective Phenomena of the Committee on National Statistics will go a long way toward charting the course of future developments in the study of response errors in attitude questions (Turner and Martin, 1981, in press). Other work promises to bring the insights of cognitive psychology into the functioning of human memory and coding abilities to bear on problems of question formulation and the understanding of respondents' answers.

---

[6] A conference on computer-assisted survey technology, sponsored by the National Science Foundation and organized by J. Merrill Shanks and Howard E. Freeman, was held in Berkeley, California, in March 1980. Proceedings should appear with the title The Emergence of Computer-Assisted Survey Research.

Nonresponse and Nonparticipation Effects

We know that those who do not answer some or all questions in a survey, who drop out of an experiment, or who are never home to an interviewer are different from those who answer, remain, or are at home--at least in terms of refusing to answer, dropping out, and being away from home. It is likely that they are different in other ways as well. And if these ways include differences in the variable(s) the study is trying to measure (e.g., income or political opinion), then the results of the survey will be biased. If, for example, the estimate of the average income of the population (or percentage in favor of a candidate) is based on data only from those who responded, it could be very different from what would have been estimated if the nonrespondents had also answered. (Recall that the estimate for the complete sample may incorporate sampling and response error.)

It is useful to distinguish between unit nonresponse and item nonresponse. In unit nonresponse, entire sets of data are missing for potential respondents because they were missed in the field (e.g., were never at home), were missed in the frame (e.g., for data being collected by telephone surveys, did not have telephones), or refused to participate. Item nonresponse occurs when an individual's answers to some parts of a survey instrument are missing or are inconsistent (e.g., wage income plus interest income plus income from other sources are greater than total income) and so are edited out in the data-cleaning process and must be replaced by a more consistent set of answers.

There is reason to believe that both item and unit nonresponse are high and getting higher, even in surveys under government sponsorship. Refusal rates for the Current Population Survey have risen from 1.8 percent in 1968 to 2.5 percent in 1976; for the Health Interview Survey (sponsored by the National Center for Health Statistics) from 1.2 percent to 2.1 percent in the same time period

(Panel on Privacy, 1979). These numbers are particularly worthy of concern when we consider that both these surveys are conducted by the U.S. government, that extensive and increasing efforts are mounted to reach respondents initially not found at home, and that each 1 percent of the American population represents more than 2 million people. The problem is not confined to the United States, however. Results of the Swedish government's Labor Force Survey show that refusals have risen from 1.2 percent in 1970 to 3.9 percent in 1977 (Dalenius, 1977).

Even the U.S. census, to which response is required by law, is not immune. In the 1970 U.S. census, data had to be imputed (filled in) by the Census Bureau for such items as age (for 4.5 percent of the respondents) and total family income (for 20.7 percent of families, although for many of these families most components of income were indeed reported) (B. A. Bailar and J. C. Bailar, in press). It is estimated that the 1970 census undercounted by 2.5 percent (or about 5 million people), even after adjusting the count whenever there was a shred of evidence to do so. (Housing and post office checks by the Census Bureau on a sample basis showed that there were some occupied buildings for which no residents were counted. These checks made it possible to adjust the count, adding some 5 million people who had not filled in census forms before the estimate of the undercount was calculated.) The problem of undercounting or nonresponse in the 1980 census has been a major source of legal challenges.

Given the conflicting pressures, it is remarkable that the census can be as accurate as it is. Many people believe that responding should be made voluntary. Nevertheless, there is both broad support and a legislative mandate for allocating funds to localities on the basis of the proportion of the residents falling into certain categories. Some of the residents in those very categories strongly prefer not to be counted, for such reasons as their receipt of illegal income or illegal immigrant status. Deciding whether indeed we want the count to be as accurate as possible or whether other values have higher priority seems to be an important issue.

Nonresponse is an even greater problem in nongovernmental surveys. For surveys having varying sponsorship, dealing with varying populations, and using varying definitions of nonresponse, one study found nonresponse to range from a low of about 5 percent to a high of about 87 percent (Panel on Privacy, 1979). If the current trend continues, the problem of nonresponse is likely to persist and even to be exacerbated. Without substantial efforts to curb nonresponse, response rates in major national data collection efforts are likely to continue to drop, so that survey results will become practically and scientifically useless. Thus the vigorous scientific activity being devoted to developing methods for reducing nonresponse, for adjusting for it when it does occur, and for properly analyzing the resulting data is crucial if we are to continue to get good-quality data from surveys.

Reasons for Nonresponse Several reasons for the rise in nonresponse have been suggested, and some have been investigated. Apathy, lack of belief in surveys, and reactions against sales pitches masquerading as surveys may well lead to refusals. Distrust of investigators and concern with privacy and confidentiality, perhaps heightened by requests for informed consent (Dalenius, 1977), may well produce both unit and item nonresponse. In an experimental survey by E. Singer (1978), a promise of confidentiality consistently decreased item nonresponse to sensitive questions. A similar experimental survey conducted by the Census Bureau under the auspices of the Panel on Privacy and Confidentiality as Factors in Survey Response of the National Research Council (1979:116) found steadily decreasing percentages of unit nonresponse (both refusals and total noninterviews) with increasing assurances of confidentiality, but the differences were small. (Perhaps the differences were small because the Census Bureau's sponsorship of the study produced relatively low nonresponse rates, regardless of the promised level of confidentiality.) We can expect more research on the causes of refusals and other nonresponse.

Nonresponse in the sense of noncoverage in the frame can be unintentionally introduced at the design stage (Morris, in press). For example, a design based on imperfectly measured variables or those that are subject to random change will exclude some part of the population. Consider a frame confined to "low-income" people. Those whose incomes in the critical year were "accidentally" higher than their permanent incomes will be excluded. (Of course, those with "accidentally" lower incomes will be mistakenly included.) Similarly, a frame that is designed to tap large concentrations of a target group will often miss atypical members of that group: for example, a frame using low-income census tracts to reach low-income people would miss low-income people living in high-income tracts.

Reducing Nonresponse Certainly the preferred method of dealing with nonresponse is to keep it from happening, although such procedures are often very expensive. Thus a battery of techniques has been developed with the general aim of encouraging the chosen respondents to participate or of systematically substituting other informants or respondents in the field.

Encouragement to respond takes many forms. In designing field operations, stress is placed on training interviewers to understand the purpose of the study and to establish rapport with respondents. Callbacks are routine (though expensive; survey organizations estimate that with a 75 percent response rate, the first 70 percent accounts for 50 percent of the cost, and the last 5 percent accounts for the other 50 percent). Especially for surveys under government auspices, enlisting the cooperation of local governmental bodies and professional organizations has proved helpful (Morris, in press). Incentives to respondents seem to be somewhat useful. (In a social experiment that used survey techniques, the Health Insurance Study, however, despite governmental backing, apparently substantial benefits, and belief in the value of the study, 19 percent of the invited households refused to participate.)

At the same time, extreme efforts to decrease nonresponse may degrade the quality of the data. Some hard-to-locate respondents can be found with extra effort, and the inclusion of their data will of course increase the response rate and probably the accuracy of the estimates. Those who refuse to participate but are pressured to do so against their will also increase the response rate--perhaps at the expense of the validity of the estimates. For example, in one study the inaccurate reporting of hospitalization by hard-core nonrespondents caused the overall estimates of hospitalization rates to be more inaccurate than if these respondents had never been questioned (U.S. National Health Survey, 1963).

Some nonresponse can be "defined away" by permitting others to answer for an individual or by substituting for respondents. In household surveys, adults are often permitted to act as informants as to the activities of other family members as well as respondents as to their own activities. While this approach is primarily a money-saving technique for reducing callbacks, it also reduces nonresponse. In the Charlotte, North Carolina, pretest of the National Health Interview Survey, for example, it was found that 50 percent more callbacks were required when each member of a family had to respond personally than when the rules were relaxed to let related adults respond for those absent (Nisselson and Woolsey, 1959). This sort of proxy reporting has been extended outside the household in network sampling.

But there is mixed evidence about the accuracy of this procedure, which may sometimes replace nonresponse with response errors. For victimization surveys, one study (Biderman et al., 1967) found that many more offenses were reported by respondents as happening to themselves than to other members of their families. Another study (Ennis, 1967) reports accurate results for white household informants but underestimates of crime rates when the method was used for black families.

Evidence for the policy implications of these strategies can be found in discrepancies in the estimates of youth unemployment, currently regarded as a major social problem. On February 29, 1980, the New York Times reported that the National Longitudinal Surveys of Labor Force Experience had found the 19.3 percent of white and 38.8 percent of black youths ages 16-21 were unemployed in spring 1979; at that time, the Bureau of Labor Statistics' figures, based on the Current Population Survey, showed the rate to be 14.1 percent for white youths and 28 percent for black youths. The report, one of the first outputs from new cohorts in the National Longitudinal Surveys, prepared for the U.S. Department of Labor by the Center for Human Research at Ohio State University, credited this difference to the fact that youths themselves were interviewed rather than other family members, such as heads of households, as is done in the Current Population Survey. It would appear that a difference in a methodological procedure increased estimates of the size of the unemployment problem among youths by about one-third.

Many surveys permit substitution, either at random from a similar group or by propinquity, for sample members who refuse or are unavailable. For example, the National Longitudinal Study, conducted by the National Center for Educational Statistics, used random substitution of schools, while the Michigan Survey of Substance Use permitted the substitution of households adjacent to the one designated in the sample. Old-fashioned quota sampling permitted interviewers to choose their own respondents as long as quotas for each sex, age group, race, etc., were filled. No probability mechanism was used. As it is currently conducted by professional pollsters (with multistage area probability sampling down to the block level and then controls on such variables as gender, age, and employment status), quota sampling can be thought of as an extension of such substitution rules. There is evidence that this "probability sampling with quotas" (Sudman, 1967) produces usable results: When the National Opinion Research Center split its sample for the 1975 and 1976 General Social Surveys between true probability methods and quotas, it found no differences between the two techniques other than a deficit of one- or two-person households in the quota samples (Stephenson, 1978).

Because assurances of confidentiality tend to increase response rates, and anonymity is the ultimate in confidentiality, many surveys routinely arrange for questionnaires to be filled out anonymously. But anonymity cannot be maintained easily in longitudinal studies requiring repeated contacts, and it is seriously compromised in personal and telephone interviews. Methods to increase confidentiality in longitudinal studies are discussed below. In telephone interviews, respondents may return calls in order to preserve anonymity, a procedure that also purportedly reduces unit nonresponse. Most special efforts to ensure confidentiality in telephone and in-person interviews deal with particularly sensitive questions, however, and are aimed at reducing item nonresponse and inaccuracy. Mailbacks of answers to specific questions have been used and in some cases the randomized response technique reduces item nonresponse (Boruch and Cecil, 1979).

Adjustment for Nonresponse Despite the best efforts of survey designers and field staff, nonresponse, both unit and item, frequently occurs and must be taken into account. What can be done then, after the fact, to adjust for appreciable nonresponse? It is logically impossible to do nothing: Simply to drop the nonresponding units from the sample is to do something, for any estimation procedures that are then implemented tacitly assume that nonrespondents are just like respondents and that the results of the survey would not have changed had they responded. Doing nothing implies a very specific but simple model: that the forces that prevented some people from responding are unrelated to the variables of interest, so that the distribution of nonrespondents on these variables is no different from the distribution of respondents. Similarly, more complex techniques for dealing with missing data also require implicit or explicit models of the causes of nonresponse and hence of the distribution of nonrespondents. Models usually assume that nonrespondents are distributed like some subset of the respondents having similar measured characteristics (covariates), but they sometimes assume that nonrespondents differ from respondents in systematic ways (as would be true if, for example, the probability that people would report their income were proportional to income).

A great number of techniques for dealing with missing data have been developed (see, e.g., J. C. Bailar, 1978; Brewer and Sarndal, in press; Kalsbeek, 1980; Little and Rubin, in press; Morris, in press). Some techniques reweight aggregations of data to take into account missing observations, and others "fill in the blanks," creating pseudo-observations in place of the missing ones. In either case the analyst must take into account that the data have been adjusted for nonresponse and that such adjustments affect estimates of the accuracy of quantities derived from the data.

A commonly used means of weighting for missing data is called ratio estimation. It uses information derived from other studies to improve estimation. Assume the quantity we wish to estimate is Y (for example, the average income for the population) and that it will be estimated by the sample mean, $\bar{y}$ (the average income for those in the sample). Assume we also know that Y is related to another variable, X (for example, the number of people per room in living quarters), for which we know both the mean for the respondents in the sample, $\bar{x}$ (mean number of people per room in the sample), and the mean for the total population, $\bar{X}$, from another source, such as the census. If we then make the additional assumption that the ratio of the mean number of people per room in the sample to the mean number of people per room in the population is the same as the corresponding ratio of mean income between the sample and the population ($\bar{x}/\bar{X} = \bar{y}/\bar{Y}$), we can use this relation to adjust $\bar{y}$ to $\bar{y}' = \bar{y}\ (\bar{X}/\bar{x})$. (Deming, 1978, presents properties of this estimator and several related ones.)

A ratio adjustment for nonresponse was used, for example, to correct for response bias in the 1975 Survey of Scientific and Technical Personnel (Tupek and Richardson, 1978). It was found that large firms were least likely to respond to the survey. The total number of employees in the firms in each size stratum was known from other sources, and the ratio of scientific and technical employees to total employees remained constant. Hence it was possible to use the ratio of the total employees in the reporting firms in the stratum to total employees in all firms in the stratum to adjust the estimated number of scientific and technical personnel in each stratum.

In the Health Interview Survey respondents are interviewed face-to-face and asked, among other questions, whether the household has a telephone. Recently investigators (Thornberry and Massey, 1978) found that health characteristics differ between households with and without telephones; they developed a ratio estimator that could be used to adjust estimates of health characteristics for the bias arising from noncoverage of households without phones, if the survey were redesigned to be done via telephone. The form of the ratio estimator should be valuable for other, similar surveys. This inquiry represents basic research into the properties of adjustment techniques and their usefulness in varying situations.[7]

Techniques that fill in missing values individually for item nonresponse are called imputation techniques. Such techniques assume that the value of the missing item can be estimated from values of other items for that respondent. One such technique uses the other items as variables in a regression function, either derived from the data at hand or available from outside sources. Such a procedure must assume (or fit) a particular functional form of the model of how the missing item depends on the other variables (covariates). For example, one might derive a formula: "Imputed income in thousands of dollars = 1/2 (age) + 5 if the respondent is male, + 2 if the respondent is white, - 3 if the respondent is both black and female - .8 x (number of people per room in respondent's residence)."

---

[7] Raking ratio estimators, a somewhat different technique (Deming and Stephan, 1940), adjust for strata much finer than the ones for which outside data are available and so must start by estimating the "outside information" for these strata. This estimation uses methods of iterative scaling also used in other sorts of analyses of cross-classified data (e.g., Bishop, Fienberg, and Holland, 1975). The estimated outside information is then used as part of a ratio adjustment for that stratum. Oh and Scheuren (1978) have offered a multivariate version of the raked ratio estimator.

In the days before high-speed computers, survey analysts often filled in the blanks caused by item nonresponse from tables put together from outside sources. Such a table might specify that if a respondent was a married white female between ages 30 and 45 who did not answer how many children she had, she should be "assigned" two children. This so-called "cold deck" procedure, of course, assigned the same number of children to all missing values for women in a specific marital status-race-age group. With the advent of high-speed computers, more flexible procedures became possible.

These "hot deck" procedures fill in the missing value for the item with the value appearing for another respondent in the same survey who is "similar" to the respondent with missing data. "Similar" is defined by the variables thought to influence the one missing (e.g., for number of children these might still be marital status, race, and age) and all respondents who are the same on these variables are said to constitute an "adjustment class" (I. Sande, in press). Hot deck procedures make no assumptions about the functional form by which the variables defining the adjustment class determine the missing item--only that they do. There are now a tremendous variety of these hot deck procedures: The simplest uses the value of the item that occurred in the previous unit processed in that adjustment class. Other variations, made possible by advances in computer science, random access, and dynamic creation of the adjustment classes, choose a donor within the adjustment class on criteria of nearness on other important variables or introduce randomness into the process of choosing a donor (G. Sande, in press a,b).

Care must be exercised when making estimates from data that have been partially imputed, because the imputation changes the estimated accuracy of the estimates. In addition, the sample size for any item is the number of respondents actually giving data for that item and should not be considered increased by the imputation.

A new idea is a process of multiple imputation (Rubin, 1978, 1979). In it the analyst repeatedly uses an imputation method to fill in missing data. Each time the complete data set is imputed, an estimate is made of the quantity of interest. One can then examine the distribution of these estimates to see if or how much they vary with different imputed data sets. If several different assumptions about the "causes" of nonresponse are plausible, a set of multiple imputations might be carried out using each assumption as the model to determine the imputation method. Then the set of sets of estimates may be compared, so that the sensitivity of the estimation to the model assumed for nonresponse is explored as well. The justification and interpretation of this multiple imputation procedure come from a Bayesian technical stance (Rubin, 1978, 1979).

One can think of multiple imputation as a program for investigating the properties of the various methods of imputation in the context of various models for nonresponse or differential response (see Heckman, 1976) using a variety of data sets. Some comparisons of the different methods have already been made (e.g., J. C. Bailar and B. A. Bailar, 1978, 1979; B. G. Cox and R. E. Folsom, 1978; Ford, 1976, 1978). So far we know that there are differences in both systematic and random error over the techniques, but no consistent pattern is yet visible.

Current Research Prospects Nonresponse, its causes, cures, methods of coping, and their properties represent active lines of research. In 1978 several sessions at the annual meeting of the American Statistical Association discussed nonresponse, which are published in Aziz and Scheuren (1978), as well as in the Proceedings volumes for that meeting. The Committee on National Statistics established a Panel on Incomplete Data, which held a symposium in August 1979. The panel is reviewing and comparing procedures used for incomplete data; summarizing theory and methods for field procedures, data processing, and estimation; and plans to make suggestions for reporting surveys so that results of nonresponse can be taken into account. (The final proceedings volume and the other volumes of the panel's work will be published by Academic Press.) Its report will undoubtedly call for more systematic studies of the performance of imputations, perhaps following Dalenius (in press) in asking for a series of simulation experiments. In such experiments complete data would be artificially subjected to nonresponse mechanisms, and analysts would attempt to estimate the (known) population characteristics and to describe the nonresponse mechanisms. As it becomes more and more obvious that the most rigorous mathematical treatment of the effects of adjustment for nonresponse is only as good as the model of the process assumed to be causing the nonresponse, it seems likely that treatments of the subject, practices, and comparisons between practices will take on a more Bayesian aspect, either formally or informally.

Total Survey Variability Revisited

Several investigations have examined the accuracy of particular surveys through the synthesizing concept of total survey error (or related ideas that examine all possible sources of variability and their impact on estimates made from the data).[8] One major study applied the concept of total survey error to the 1970 national health survey of the Center for Health Administration Studies and the National Opinion Research Center, which collected data on health services use and expenditures (R. Andersen et al., 1979). Verification data were collected from health care providers (doctors, hospitals, etc.) to compare with respondents' reports in order to measure response errors. The effects of nonresponse and of different approaches to the imputation of nonresponse were also investigated. One important finding of the study was that conclusions about the differences in health care experiences between important subgroups of the population (the elderly versus others; the poor versus others, etc.) changed very infrequently when adjustment was made for those parts of nonsampling error that could be measured. The magnitude of the differences, however, changed more often. The verification process was a lengthy and expensive one (18 months, accounting for about one-third of the $1 million cost of the survey), so whether it should be incorporated more regularly into surveys depends on the anticipated changes in estimates that adjustment for error will cause. Probably several more such large-scale efforts in different fields of application will be necessary before such anticipations can be made with any degree of confidence.

An error profile has been compiled for the measurement of employment by the Current Population Survey (Brooks and Bailar, 1978). Such a profile is related to the concept of total survey error but is constructed by following the operations of a survey step by step, from the construction of the sampling frame through the publication of results, pointing out possible sources of all kinds of error and presenting evidence of their direction and size when such estimates are available. This effort, the work of the Subcommittee on Nonsampling Errors of the Federal Committee on Statistical Methodology, was intended to serve as a model for such profiles for other major governmental surveys and as a first attempt deliberately chose not to consider such matters as conceptual errors; these matters will probably be addressed in future profiles. Another error profile has been compiled for multiple frame surveys by Norman Beller at the U.S. Department of Agriculture. It would seem that this sort of project, while also expensive and time-consuming, will point to gaps in knowledge about nonsampling errors and stimulate efforts to fill the gaps.

---

[8] Other approaches to the modeling of measurement error than that implied by the concept of total survey error are of course possible and have been suggested. For example, an approach using a set of structural equations to model the relations between a group of questions that pertain to the same underlying concept, each measuring it imperfectly but together capturing most of its richness, is now widely used. (See Jöreskog and Sörbom, 1979; for a clear exposition of this approach and an enlightening application, see Kohn and Schooler, 1978.)

The Office of Energy Information Validation of the U.S. Department of Energy was created specifically to understand the error structure of data collection and analysis in this important policy area and to improve practice. Work in that office: (1) investigated information needs and whether they are being met by data collection systems, proposing improvement when necessary; (2) carried out studies aimed at validating data, looking at the effects of response errors and of nonresponse and imputation; (3) studied the workings of models used to make estimates and predictions from the data; and (4) examined publications of these estimates and predictions for their informativeness, ease of comprehension, and clarity in explaining the meaning of estimates and the amount of uncertainty they are likely to contain. A self-conscious effort to document procedures used to accomplish these tasks was also undertaken.

The Research Triangle Institute is at work on a taxonomy of errors as an early step toward the institution of a survey design information system. Such a system (Horvitz, 1980) would store information about specific variables as they have been measured in social surveys, including the context of the survey, sample design, the wording of questions, error components, and costs. This is a concept even broader than that of total survey error and should serve to standardize survey measures, integrate knowledge of survey error components, improve survey design, and provide a broad base for methodological research.

LONGITUDINAL SURVEYS

How many Americans are poor? The answer depends on what one means by the question. According to the Panel Study of Income Dynamics, in a single year (1975) 9 percent of the American people were below the official poverty line, 25 percent were below it in at least 1 of the 9 years before 1976, but only 1 percent remained in poverty for the entire 9 years (Morgan, 1977). These distinctions can make a difference. For example, strategies for effectively assisting the chronically poor are probably very different from those most effective in aiding the temporarily poor; decisions about the magnitudes of the efforts would probably depend on the relative sizes of the two groups.

For our purposes, the crucial point about these differing figures is that they could have been found out (without unduly trusting people's memories) only by questioning the same people repeatedly--that is, by a longitudinal (or panel) study rather than a cross-sectional one. A cross-sectional study could have estimated the number of poor only for the year of the study, giving no data (except those based on fallible memory) on the number of the persistently or occasionally poor. The difference is like that between a snapshot of a crowd, which enables us to make some aggregate measure, such as the number of people present, and a motion picture, in which we can see the aggregate size of the crowd at each moment and also follow the activities of individuals as they leave or enter the crowd over time. The implementation of such large-scale longitudinal studies gained impetus in the early 1960s, corresponding to the start of large-scale social experiments (Kalachek, 1979).

Advantages of Longitudinal Studies

The distinctive feature of a longitudinal study is that it permits an investigator to follow people (or other individual units of analysis, such as families or organizations) over time; this means that data on individual changes, rather than only aggregated movements, are available for analysis. Thus research can focus on process by asking how and why and often for whom such changes occur. For example, in studying life-cycle processes a panel study might address such a question as "Does early unemployment among teenagers and youth represent a transitory phase that many go through with no particular long-run adverse consequences, or does such a period of unemployment lower future earnings and/or increase the proportion of time in later life that an individual is unemployed?"

The aftermath of the 1980 presidential election provides an example of the use of panel data to illuminate process. A CBS/New York Times poll questioned a large national sample during the week before the election and was able to recontact 89 percent of the respondents who were registered voters in the few days following the election. The gap between the two major candidates increased by about 7 percentage points in the week between polls. But this is aggregate or "net" change. It was actually brought about by some 21 percent of the registered voters polled who changed their minds (some from Carter to Reagan, some from Reagan to Carter, some from Carter to deciding not to vote, and so forth). Respondents who reported votes different from the ones they anticipated before Election Day were asked for reasons for the change. Thus the poll was able to conclude that "news of the Iranian conditions for releasing the American hostages that broke the Sunday before Election Day was a major element in those shifts, . . . but so, apparently, were last minute rejections of Mr. Carter's handling of the economy (New York Times, November 16, 1980:1).

The existence of large-scale longitudinal data sets has inspired both methodological and substantive research and has drawn attention to the need for developing new methodological tools for their analyses. One example is new applications of mathematical models. These include statistical models that treat time as continuous and thus are more likely to coincide with our theoretical understanding of social processes and more likely to represent faithfully actual behavior than are models that treat time as discrete. People do not change jobs, break up marriages, or accomplish any of a myriad of other status changes at specific (discrete) times (such as the time they are asked about their status on these variables). Thus any decision about the proper length of the time chunk to consider is necessarily arbitrary. A weekly survey is probably frequent enough to observe whether job changes occur--but is a monthly one frequent enough? And analyses that make these arbitrary decisions differently for use in discrete time models can produce substantively different results. Further, continuous time models are often computationally simpler than discrete time models.

In addition, most human behavior is more complicated than the simplest models must assume. One discrete time model of employment, for example, would define three "states": employed, unemployed, and out of the labor force. It would then need to assume that the chance of a person's being in a particular state in the next time period (e.g., month) depends only on which state that person is in during the current time period. Past history, including the amount of time in the current state, is taken to be irrelevant. (This is called a Markov model.) Clearly the world is more complicated than that. Some people are more likely to stay in the same state from month to month than are others. For example, unemployed members of a particular ethnic group may be more likely than members of other groups to remain unemployed once they become unemployed. This is the "mover-stayer" model, which has inspired a good deal of work in studies of social mobility (see Pullum, 1978, for a review).

Another model would assume that the chance of a person's changing states depends on how long he or she is in that state--the longer one has been unemployed, the more likely that one will remain unemployed, for example. Or the chance of moving from one state to another depends not only on one's current state but also on one's prior history; a history of moving continuously into and out of the labor force may suggest that one is more likely to move out of the labor force next month, even though one is currently working, than someone else who is currently working but has not been out of the labor force since high school graduation. Various combinations of these assumptions may also apply.

Each of these verbal descriptions of the world implies a mathematical model. The availability of longitudinal data makes it possible to test which model presents the most accurate picture of the world, as it is and as policies would have to cope with it. If data are really continuous, constituting a life history for each individual, then both the choice of the proper statistical model and the estimation of its parameters are more easily accomplished than if the data are fragmentary, available only at some points in time (B. Singer and S. Spilerman, 1976). It is always important, prior to data collection, to consider what models will be fitted, because data irrelevant for one model may be crucial for others. For example, do we want to measure current state only? Duration in current state? Number of switches in state during the period between interviews?

Research is needed on design for panel studies to facilitate discrimination among models fitted to the same fragmentary data. Such research should address questions of the optimal spacing between interviews to balance problems of reliability of retrospection versus costs and delays of reinterviewing (B. Singer and S. Spilerman, 1976).[9]

An application of a continuous time model has come out of the longitudinal data generated by the income maintenance experiments (Tuma et al., 1979). Three models (one independent of time, one contrasting the first 6 months of the experiment with the succeeding 18 months, and one looking at four successive 6-month periods) were used to investigate the impact of support levels on attrition (or withdrawal) from the experiment and on marital dissolution and remarriage. The findings showed no effect of support level on attrition (cheeringly enough), no systematic effect on remarriage, but a systematic effect of support level on marital dissolution. (Marriages of women receiving income supplements were considerably more likely to break up than marriages of control women; the effect was most noticeable during the first six months but continued throughout the 2-year period.) The model contrasting the first 6 months with the succeeding 18 closely predicted the percentage of the sample single at each time over the 2-year period, suggesting that the two-period model embodied a process of marital dissolution that is compatible with the data. (Another less substantive, but beautifully explicated, application of a continuous time model appears in B. Singer and S. Spilerman, 1977.)

Organizing Data Longitudinally

The organization of data collected longitudinally presents many challenges; as researchers meet them we shall see both methodological progress and rich substantive results. Many data sets that are collected longitudinally are stored in computer files as if they were merely cross-sectional, so that many of the special benefits of longitudinal data cannot be realized. Moreover, the analytic richness of longitudinal data is unavailable without cross-referencing between levels of aggregation. Each of these challenges is discussed in turn.

---

[9] The research on the relationship between forgetting and telescoping (Sudman and Bradburn, 1973) is relevant here, as is consideration of the advantages of irregularly spaced interviews.

The first challenge of organizing data from longitudinal surveys arises because different numbers of events happen to different people. One person may be hired and fired many times over the years, generating data on the job description and dates of employment for each job. These data must be stored and catalogued as pertaining to this particular person. Another person may stay in the same job throughout the course of the longitudinal study, generating far less data. It becomes a methodological challenge to arrange a computer file that includes all the data for all respondents.

The simple solution allots each respondent the space necessary to record the data for the respondent with the most job changes. This creates an easily used "rectangular" file but uses a great deal of computer space suboptimally and increases the time necessary to access any piece of data. Another strategy is to use a hierarchical, nonrectangular file structure. This economizes on computer space and access time but creates the need for new computational and statistical methods. Such issues of file organization and their consequences constitute an active research area (see, for example, Ramsoy and Clausen, 1977); nevertheless, data files are already beginning to become available in longitudinal form.[10]

Another challenge for file organization of longitudinal data arises from the need to link various levels of analysis. Often a survey is conducted so that locations (for example, housing units) are sampled. Within the housing units are households, made up of individuals. Typically a separate computer file is maintained for locations, for households, and perhaps for individuals. In the case of an individual who has experienced an event (a robbery, for example), we may want to ask several levels of questions. Had that person previously been robbed? Has anyone else in the same household been robbed? Was any member of the household that previously lived at this location robbed? These questions are answerable only if efficient cross-referencing between the several data files has been provided.

---

[10] For example, the National Longitudinal Surveys of Labor Force Experience sponsored by the U.S. Department of Labor (Bielby et al., 1977), the Annual Housing Survey sponsored by the U.S. Department of Housing and Urban Development (Beveridge and Taylor, 1980), the Panel Study of Income Dynamics (Morgan, 1977), and some parts of the National Crime Survey sponsored by the Law Enforcement Assistance Agency (Reiss, 1980) can now be used longitudinally.

Resources of Longitudinal Data

Longitudinal data today represent an underutilized resource. We are just beginning to explore the richness of the data sets that have been deliberately collected longitudinally. But there are other data sets, which are only fortuitously longitudinal, that represent an even less exploited resource. The National Crime Survey, which asks respondents to report victimization, and the Current Population Survey are both in part longitudinal data sets (see Fienberg, 1980a; Kalachek, 1979). For reasons of economy in sample selection and the control of certain kinds of bias, each of those surveys uses rotating panels. The Current Population Survey interviews each family eight times: once a month for four months and, after eight months off the panel, once a month for four months again. The National Crime Survey also interviews monthly, with a rotation group being interviewed every six months for three years. Each of these samples is designed to give cross-sectional data. That is, the Bureau of the Census, which runs both of these surveys, is interested in aggregate employment and unemployment figures each month and in aggregate victimization each month.

With some effort, however, the surveys could be organized in longitudinal form and used to examine changes experienced by individuals. Some attempts in this direction have been carried out. A longitudinal data file for persons and households present in the National Crime Survey from July 1, 1972, to December 31, 1975, has been created (Reiss, 1980). This file has been used to investigate repeated victimization using log linear models (Fienberg, 1980b). (Note that the analysis of any repeated event is inherently longitudinal.) Because repeated victimization frequently involves crimes of a similar type, further investigation might examine the vulnerability or "proneness" of groups of households or household locations to certain kinds of crime.

There are both limitations and advantages to using the Current Population Survey as a longitudinal survey

(Kalachek, 1979). Its advantage is that it is enormously large--56,000 households are interviewed each month, in such a way that 42,000 are common to successive months, and 28,000 common to the same month across a year. This is in contrast, for example, to the original panels of the National Longitudinal Surveys (the Parnes survey), each of which contains 5,000 individuals. The breadth of the Current Population Survey would permit analysis by subgroups; this is not feasible for the smaller panels. The Current Population Survey, for example, could examine the employment experience of black women from the South in a particular age group, while the National Longitudinal Surveys would have too few people in such a specific category to carry out those analyses.

In other senses, however, the Current Population Survey is limited. The length of time any given family is included is only 16 months. Furthermore, in order to serve its primary purpose--the monthly collection of timely cross-sectional unemployment statistics--the Current Population Survey interview schedule must be kept brief in order to reduce nonresponse. Thus, the in-depth data available from special panel studies are not available from it.

Some of this lack of depth could be compensated for by supplementary questions to the Current Population Survey that are asked once a year. (Thus each family would give two readings on each of these questions, spaced a year apart.) The supplementary questions encompass such areas as job tenure and job mobility, marital and family characteristics, education and work experience, multiple job holding and union membership, school attendance, and farm labor. The Current Population Survey data files could also be supplemented by statistical matching, described below. A large number of important policy issues could be better examined if these potentially useful files were made available for analysis in longitudinal form.

Other Aspects of Longitudinal Studies

Besides the challenges of the organization of data files, longitudinal studies present design and analytic problems as well. Evidence of what is called "panel bias" suggests that people answer questions differently the second time (and perhaps subsequent times) they are asked than they do the first. Perhaps some purchase can be gotten on this problem by "throwing away" the first interview with a respondent. The National Crime Survey, for example, uses the first interview for bounding purposes only, not for comparative purposes or as cross-sectional data. Thus the first "real" interview (the second actual interview) is more like subsequent interviews than it is like the first. The severity of panel biases, their effect on measures of change, and the extent to which they continue over time in a panel will be matters for investigation as data become more easily available in longitudinal form.

A second problem is that when a family takes part in a survey over time, different family members may be interviewed on different occasions. In the discussion of using proxy respondents to decrease nonresponse, we noted that some respondents report differently about themselves than about other family members. Do such differences occur in longitudinal studies, and, if so, what effect do they have on data analyzed longitudinally in terms of families? In terms of individuals?

Still another set of problems with longitudinal surveys involves attrition from the sample. If the housing unit is the sampling unit but the family is the unit of analysis, what happens when the family moves? What happens when part of the family moves, as when a grown child leaves home or a marriage breaks up? What happens when a person dies? This problem has received little attention from the Census Bureau because, for their cross-sectional purposes, the household is treated as the unit of analysis. Other large-scale longitudinal studies have answered these questions in various ways. The Health Insurance Study replaces families that move out of the area and hence become ineligible to participate with those who move into the vacated dwelling. In cases of the divorce and remarriage of both spouses, the Health Insurance Study chooses one spouse and follows the new family, dropping the other spouse from the sample. The Panel Study of Income Dynamics follows all members of the families originally interviewed in 1968, annually interviewing the head of every family that includes at least one member of the original families. Thus the sample keeps renewing itself with new generations.

Following individuals over a long span of years can be particularly difficult, especially if there is a considerable hiatus between interviews, but the success of recent studies suggests that it can be accomplished. The secret seems to be to tell the respondents that the study is a continuing one and to ask them to give the names of several relatives or friends who would always know how to reach them (Freedman et al., 1980). This seems to be a good idea for studies dealing with such wholesome activities as family building and career planning; problems of confidentiality might well arise if the issues were more sensitive.

What happens when a heretofore cooperative respondent disappears or refuses to answer some or all questions? The longitudinal nature of the studies itself helps in the solution of such problems. Certainly imputation for item nonresponse (unanswered questions) can be more easily accomplished in longitudinal studies, in which there is prior information about respondents. For example, estimating a respondent's income this month is easier if we know his or her last month's income. Several investigators have presented models that help deal with attrition from longitudinal surveys. They first model the probability of attrition (or nonresponse or self-selection) based on respondent characteristics that are measured within the context of the survey. For example, a model might suggest that the likelihood of moving, and hence being unavailable for interviewing, increases with the experience of having been the victim of a crime. Researchers can then attempt to adjust estimates of current victimization or current unemployment, for example, for bias caused by attrition (Griliches et al., 1977; Hausman and Wise, 1977; Heckman, 1976, 1979).

This problem of attrition bias is a special case of the more general problem of "censoring." A respondent who leaves the panel cannot, of course, supply data thereafter; such further data are said to be censored. But even with an intact panel of willing respondents, some data are not available at any given time for some individuals. To illustrate: At whatever time we stop collecting data to make an analysis of the amount of time spent in a first job, there are some people who have never switched jobs; for them, we can get no measure of how long the first job lasted except to say that it lasted at least from the beginning of the job until the current time. The problem is how to adjust for the censored observations in estimating the average time spent by members of the population in their first job. There has been a recent surge in the development of methods for the analysis of such censored data. (A basic reference is D. R. Cox, 1972; a recent brief review of the literature is Moses, 1978; and detailed technical expositions are given by Kalbfleisch and Prentice, 1980, and Elandt-Johnson and Johnson, 1980.)

Problems of anonymity and confidentiality are especially severe in longitudinal studies, because individuals or families must be identified in some manner so that they can be followed over time. Several means of ensuring anonymity in longitudinal surveys have been developed, however, (Boruch and Cecil, 1979): Respondents may choose aliases and continue to use them; an agency or broker may act as an intermediary between respondent and investigator, releasing only unidentified data to the investigator; or an insulated "link file" system may be created. In this last case, data in the investigator's files are labeled by arbitrary data-linking numbers, identifications are kept in another file and labeled with another set of arbitrary respondent-identifying numbers, and the only file linking the two sets of arbitrary numbers is held by an incorruptible third party. As successive waves of data arrive, the investigator removes the identifications and relabels with the respondent-identifying set of arbitrary numbers and ships the data to the third party, who removes the respondent-identifying set of arbitrary numbers and substitutes the data-linking set before returning the data to the investigator--ponderous, but seemingly foolproof, and well-adapted to reducing both unit nonresponse and possibilities of breach of confidentiality in longitudinal surveys. Organizations such as the National Opinion Research Center have used link file systems and find that it is crucial yet understandably difficult to convince potential respondents of the inviolability of the linkage system.

SOCIAL EXPERIMENTATION

Taking experiments out of the laboratory and into the field is not new; what is new is using them as instruments of policy and simultaneously as sources of information about policy. There has been a flowering of experimentation to investigate policy alternatives and a corresponding blooming of the the methodology to carry out such experiments since the early 1960s. A standard definition of a social experiment states (Riecken and Boruch, 1974:3, emphasis in the original):

> By experiment is meant that one or more treatments (programs) are administered to some set of persons (or other units) drawn at random from a specified

> population; and that observations (or measurements) are made to learn how (or how much) some relevant aspect of their behavior following treatment differs from like behavior on the part of an untreated or control group also drawn at random from the same population.

The term random in this definition is the hallmark of a true social experiment. If people are assigned to treatments randomly (rather than by some other method such as self-selection, first-come-first-served, or those judged to be most in need given preference, etc.), then several advantages accrue. First, standard procedures of statistical inference are appropriate for use. Second, any differences found between groups at the end of the experiment can be probabilistically examined to judge whether they result from the treatments or reflect instead preexisting differences between the groups related to whether or not treatment was received.

Controversy about the necessity and feasibility of doing such randomization has long existed. That randomization is important is clear: When randomized and nonrandomized evaluations of programs (such as the Salk polio vaccine tests) are run in tandem, the result of the nonrandomized study is often less clear and compelling than the result of the randomized study (Boruch, 1975; see also Gilbert, Light, and Mosteller, 1975).

Thus it is not true that nonrandomized testing is cheaper and just as good as randomized experimentation. (This and other contentions about the infeasibility of random experiments are listed and refuted by Boruch, 1975.) Nevertheless, while true experimentation is the method of choice for drawing conclusions about public policies and programs (as well as about other issues) and can be carried out more frequently than it currently is or than is often supposed, inferences must sometimes be made from nonexperimental situations. That need can arise (among other reasons) from the pressures of time, expense, or ethics. When circumstances demand that investigators make do with data gathered from nonexperimental situations, prudence insists that special care be taken in the analysis[11] and in making inferential claims. Causal inferences are shaky at best, and alternative explanations for results are always conceivable.

---

[11] Analytic techniques have been explicated recently by Reichardt (1979) and Anderson et al. (1980).

The definition of a social experiment stresses that people are drawn at random (via a probability sample) *from a specified population*--this is an aspect of social experiments that distinguishes them from laboratory experiments as usually performed in the social sciences, for which typical subjects are rats or college sophomores. If the purpose of a social experiment is to find out how poor people will react to an income subsidy, then the subsidy must be offered experimentally to poor people. This means that the sampling techniques developed by survey researchers and statisticians are relevant to social experiments, as are the techniques for reducing or coping with nonresponse. Furthermore, the observations, measurements, responses, or outcomes one wishes to examine are usually not as clear-cut as they are in laboratory experiments, where test scores or behavior counts (often computerized) are fairly straightforward. Social experiments have sought to measure program effects such as earned income, the demand for housing, the utilization of health care services, and the distribution of electrical power use. These complicated concepts are often measured using survey techniques, so the problems faced and the knowledge gained in the study of nonsampling variability in nonexperimental surveys are equally applicable to social experiments.

Approaches to the Special Features of Social Experiments

Both the advantages and the special problems of social experiments stem from their scale--that is, their length and complexity. Besides the basic advantage of precision of causal inference, social experiments generate rich data sets that provide opportunity for detailed analysis and model fitting. When many variables are measured over long time spans, unanticipated results can be explored. (Recall the serendipitous finding that income maintenance increased the rate of marital dissolution of the supported groups.)

Social experiments take a long time to run. Families do not usually react instantaneously to an income supplement or other major program change. The income maintenance experiments provided support for three- or five-year periods, with one variation in Denver actually scheduled for 20 years (Ferber and Hirsch, 1979). An experiment monitoring the same people over this long a period falls into the category of longitudinal study and thus suffers from the problems of these studies (e.g., attrition and the censoring of data) as well as other problems related to its large scale.

Attrition can be especially damaging in an experiment if it is related to the treatment given. For example, members of a control group or one receiving only minimal benefits from the experimental program may be more likely to drop out than those receiving high benefits. (The entire control sample in one location of the Health Insurance Study was dropped because of lack of cooperation; see Ferber and Hirsch, 1979). Depending on how this attrition is related to the treatments and characteristics of the participants, it may present problems in gauging the effects of the treatments.

Attrition and nonresponse are not the only reasons for missing data in such studies; some variables are intrinsically unobservable for some participants. For example, time until rearrest among parolees is not measurable for those who are not rearrested during the course of the experiment. Again, threshold models (e.g., Heckman, 1976, 1979) may be useful in adjusting for such missing data in an experimental situation. This sort of adjustment was done in an experiment examining employment and earned wages of ex-convicts. The treatments were varying levels of income supplements (Ray et al., 1980). These income supplements were estimated to have a greater positive impact on wages when the data were adjusted to account for those who were not working.

Social experiments also require a great deal of time to plan and manage. An experiment involving trial work periods for recipients of social security disability insurance is an example. As of summer 1979, a team of researchers had spent two years in planning, setting budgets, designing measurement devices, and negotiating legislative authorization. They anticipated another year of "facing the contracting process for outside data collection services and working with Social Security operational components in mounting and monitoring the experiments" (Franklin, 1979). These experiments, the largest social experiments ever undertaken, involving some 30,000 people, have now been authorized by the sponsoring agency.

As a potential buyer examines a horse's teeth to validate the seller's claim of its age, so is a social experiment designed to look into the horse's mouth and, by implementing the proposed program on a controlled and relatively small scale, estimate the effects of a fully implemented program. There is a set of problems, however, in that the horse examined experimentally may not be precisely the same as the one that would exist if the program were fully implemented. These potential sources of bias are recognized by experimenters working in the field, and efforts are being made to measure and control for them (Ferber and Hirsch, 1979). The first such bias--attrition bias--has already been discussed.

Second, people often behave differently when they are in an experiment, simply because they know that they are participating in one. This is the well-known Hawthorne effect.[12] Several social experiments have addressed this problem by including a control group that is not measured at all until the experiment is either over or well under way; the people in this group thus do not know they are in the experiment and their data at the end of the experiment provide a baseline for the measurement of possible Hawthorne effects. In the Health Insurance Study, for example, some members of the control group did not receive an initial physical examination until six months into the experiment.

Next, community effects arise if people's behaviors are conditioned by social norms that would not apply if the program were implemented fully rather than experimentally. For example, a work ethic may operate to keep people in the labor force during an income maintenance experiment, but it may cease to operate if income supplements are instituted throughout the community. Some progress in measuring community effects may come from the housing allowance experiments. In them one component, designed to measure the impact of housing allowances on the supply of housing, offers the program to all eligible families in a housing market rather than to a random sample of such families, as is done in another component of the experiment designed to measure demand. Differences in the behavior of the families in the two component experiments may give an estimate of community effects.

Finally, when families know that the experimental program will last only for a specified period of time, time horizon effects may influence them to behave differently than they would if they knew that a program were permanently in place. Attempts to measure these effects involve varying the length of time that the experimental program will run as one component of the treatments to which participants are randomly assigned. The income maintenance experiments varied in time, usually from 3 to 5 years, and some participants were even assigned to a 20-year treatment. Clearly, policy makers will not wait 20 years for the results of the experiment, but the behavior of the 20-year group during the early years of the experiment ought to give some clues to the behavior of those for whom time horizon effects do not operate because they expect the program to be "permanent." Both these time horizon effects and start-up effects (people reacting rapidly, for example, to the treatment of a health insurance experiment by undertaking medical care that had been long neglected) lend themselves to analysis by the statistical models discussed in connection with the advantages of longitudinal data.

---

[12] Recent reexamination of the data from the Hawthorne experiments suggests explanations other than the eponymous effect account for the startling rise in productivity of the workers in the experimental room with every change in working conditions, even deleterious ones. Even if the Hawthorne effect did not occur at the Hawthorne factory, however, it has certainly occurred elsewhere.

Large-scale social experiments present management problems to investigators. Not only are data collection procedures complicated by the enormous number of variables to be measured and kept track of during the long time period of the study, but the actual delivery of the treatments is also a very complicated matter, often involving the experimental constituting of a complete social welfare agency. Thus problems of experimentation come to include the coordination between the organization delivering the treatments and the organization measuring their effects. (Several sets of practical advice have appeared, e.g., Archibald and Newhouse, 1980; Riecken and Boruch, 1974.)

From these complicated management problems of social experiments arise the twin problems of assessing treatment strength and assessing treatment integrity. In fact, whenever an experiment is conducted treatment integrity is problematic--one may ask whether or not the experimental treatments that are formally prescribed for subjects are actually given to or experienced by them. There is ample evidence from the literature of social psychological experiments that often they are not. Experimenter effects indicate that experimenters tend to get the results they expect, while in the same experiment, workers with contrary expectations, putatively using the same techniques, get contrary results (see, e.g., Rosenthal, 1966). Evidence on demand characteristics shows that subjects, trying to be cooperative, take cues from the experimental situation to develop hypotheses about what the experiment is trying to prove and then behave in a way they think will help prove these hypotheses (see, e.g., Orne, 1962). Subjects in the social role of the experimental subject tend to behave in a way that will, in some sense "make a good impression" (see, e.g., C. N. Alexander and G. W. Knight, 1971).

These problems are exacerbated in large-scale social experiments, simply because their scale is so much larger. The treatments are more complicated to explain, both to those administering them and to those receiving them--and indeed may change operationally over time (actual cash transfers change from month to month in the income maintenance experiments, depending on earnings and the "tax rate"; payments in the Health Insurance Study change depending on whether the "deductible" has been fulfilled). Responses are also likely to change according to the participants' understanding of the treatment rather than the experimenter's intentions. For example, using respondents' understandings of the treatments in the negative income tax experiment produced estimates of the program's effect that were different from estimates based on what the "actual" treatments were supposed to be (Nicholson and Wright, 1977).

On a less psychological plane, the issues of treatment strength and integrity are intertwined. Strength of treatment involves whether an intervention is powerful enough to be expected to have some effect if it is applied as prescribed. Treatment integrity involves whether the treatment applied approaches its prescribed strength, basically asking "What really happened?" Both issues should be faced in the design and analysis of social experiments, though both are difficult and neither is currently investigated routinely (Sechrest and Redner, 1979). One literature review found that of 236 evaluation studies examined, 22 percent did not measure at all whether the program had been implemented according to stated guides (Bernstein and Freeman, 1975).

Both issues have been addressed in the area of criminal rehabilitation. For example, the question of how much work release might be effective subsumes such question as: "When should work release begin?" "How long should it last?" "How good a job at what pay level is required?" "Will it be effective if it is in a community other than the one to which the prisoner is likely to return?" Similar "how much?" questions can be asked about other rehabilitation efforts, such as job training and group counseling (Sechrest and Redner, 1979). While there are no easy answers to these questions, the very act of asking them brings common sense and theoretical as well as empirical knowledge to bear on their possible solutions (Sechrest et al., 1979).

The question of treatment integrity--how much treatment was actually delivered--has been referred to as the third face of evaluation. (The first two faces are experimental design and measurement of outcome; Quay, 1977.) In a setting of group counseling in a prison, questions about what was done might ask how often the groups really met, how good were attendance and participation, how meaningful was the discussion, how well were the leaders trained and motivated, and how much stability the groups maintained. When such questions were asked in one study (Quay, 1977), the answer to each was "not very." Progress, then, in methods for conducting experiments as well as in substantive knowledge is likely to be achieved by routine scrutiny of experiments beforehand for expected treatment strength and afterward for treatment integrity.

Optimal Design

Social experiments are expensive in time, effort, and money. Including both transfer payments and costs of research, the income maintenance experiments had cost about $70 million by 1975, the Health Insurance Study was projected to cost approximately $50 million, and the housing allowance experiments about $200 million (Ferber and Hirsch, 1979). For this reason it becomes crucially important to design them in such a way that the most knowledge possible is gained from a given expenditure. Methods of optimal experimental design are engineered to do just that (see Aigner, 1979, for an excellent discussion). Let us look at the income maintenance experiments as an example. Sets of participating families were given different treatments in the experiments. These treatments included varying the support level of payments, defined as percentage of the poverty level, and varying the "tax rates," the percentage of earnings that was counted (taxed) against support payments. The response measured was earnings during the experiment, expressed as a proportion of normal preexperimental earnings. Also used as part of the basis for assignment to treatments was the preexperimental income level expressed as a percentage of the poverty level (Conlisk and Watts, 1969).

Traditional experimental design (see, e.g., Cochran, 1978) would have chosen a set of support levels and tax rates, perhaps basing that choice on considerations of what differences between combinations would be of greatest policy importance. Traditional experimental design would have randomly divided the families in a preexperimental income stratum among the treatment groups, doing the same with each additional stratum of families. The experimenter might well decide that certain treatment combinations would receive no families from certain strata or that certain treatment combinations would receive a disproportionately large number of families. Such decisions would be based on deliberate judgments about which higher-order combinations of treatments and strata would be negligible in their effects on earned income and which would be especially important.

This is a very general and flexible design that makes no assumptions at all about the form of the relationship between the response and the treatments, nor, in its simplest form, about whether families that differ in preexperimental income will differ in their response to the treatments. It also does not take into account that some treatments (with higher support payments and lower tax rates) are more expensive than others, especially when given to families with low preexperimental incomes.[13]

If one is willing to make some assumptions about the form of the relationship between treatments and response, statisticians studying response surface methodology[14] have shown that one can choose treatment combinations that are "optimal"--that will in some sense maximize the information available from the experiment for a given budget. Alternatively, if the functional form is assumed and the logically possible treatment combinations specified, optimal allocation of families to treatments can be worked out (Aigner, 1979). This latter route was taken by Conlisk and Watts (1969) in designing the New Jersey negative income tax experiment, using an allocation model that took into account both the cost of the treatment and its importance to policy.

---

[13] The flexibility of more complicated traditional experimental design is well illustrated in a redesign proposed for the Kansas City Preventive Patrol Experiment. Redesign would allow evaluation of treatment integrity that might be jeopardized when patrol cars from beats with experimentally increased manpower crossed to beats in which manpower remained the same (see Fienberg, Larntz, and Reiss, 1976).

[14] Bibliographies are supplied by W. J. Hill and W. G. Hunter (1966) and Herzberg and Cox (1969).

One problem with such optimal designs is that they are usually optimal for the form of the relationship specified; if that specification is wrong, then the design can be less good than one that makes fewer assumptions. This problem can be more easily described in the context of a simple experiment involving only support payments as the treatment and earned income as the response. If the relationship between support and earned income were believed to be linear, then some fraction (one half, other things being equal) of the available families would be randomly assigned to the lowest support level contemplated and the remainder to the highest. This procedure would be the most efficient for estimating the slope and intercept that fully describe the assumed linear relationship between support and income. If, however, the relationship between support and income is actually curvilinear (for example, if earned income decreases very slowly for low support levels but then drops off rapidly at higher levels), this design would be completely unable to describe the form of the curve. A design that assigned an intermediate support level to some families would be necessary to describe the curvilinear relationship.

This sort of problem may indeed have arisen in the New Jersey negative income tax experiment because few very poor people were assigned to the expensive treatments (because expensive treatments would have been even more expensive if applied to low-income people). It was therefore difficult to test the truth of the assumption that the effect of the treatments on labor force participation did not vary with preexperimental income level (see Archibald and Newhouse, 1980). An additional problem is that any large experiment has many questions and goals. What is optimal for one goal may be less than optimal for another.

Another approach, a "finite selection model," starts out with the number of families to be allocated to each treatment already decided and assigns, from the pool of available families, the most appropriate families for each treatment. Randomization can be easily introduced, and the computational costs of the assignment process can be reduced considerably without seriously compromising optimality, by making each choice as the optimal one from a randomly chosen subset of the available families. Some problems introduced by the possibly unbalanced designs arising from the Conlisk-Watts procedure are avoided in the finite selection model. The model was developed for the Health Insurance Study and used in it and elsewhere (Morris, 1975).

Halfway between the broad general-purpose traditional designs that assume little or nothing about the response functions and the tightly specified optimal designs is the concept of evolutionary operations (Box, 1978; Madansky, 1980). A small set of treatments is chosen and enough observations are made to suggest whether a simple model will fit the data as well as what changes ought to be made in the treatments in order to maximize (or minimize) the response. (In the New Jersey negative income tax experiment the aim would clearly have been to find the treatments that maximized earned income.) Another small experiment would then be run, either taking more observations in the original treatment region, in order to fit a more complicated model, or changing the treatments, in order to move toward the maximum response.

Any serious attempt to apply evolutionary operations to social experiments would certainly further increase the already long time periods necessary to obtain definitive results. Each of the small experiments would have to run for some time before its outcome could be determined for use in planning the next small experiment. But such an approach would be very useful in estimating the real responses to possible variations in policies in the long run, while offering some interim results that could be useful in more immediate policy planning.

A similarly sequential approach has been used to explore educational alternatives for underachieving Hawaiian native children. Compared with evolutionary operations as applied to large-scale social experiments, this project is much smaller in scale, uses less formal and complicated statistical technology, and admits more intuitive elements into its evidential base. But the basic idea of letting data from one phase of the investigation formally shape the succeeding phase is similar, and the project has developed an educational program that appears to work (Tharp and Gallimore, 1979). Perhaps the next few years will see an attempt to apply real evolutionary operations methodology to large-scale social experiments.

USES OF ADMINISTRATIVE RECORDS

As information demands become greater, it is natural to look for sources of data that involve less burden on respondents and lower collection costs than those incurred by surveys or experiments. The large sets of administrative records maintained by federal agencies (for example, the Social Security Administration, the Internal Revenue Service, etc.) seem to offer enormous potential for statistical uses.[15]

Linkage

Advantages could be realized by using individual administrative data sets themselves. More might well be available if data on individuals or businesses held by different administrative agencies or obtained by surveys could be linked together to form a file containing considerably more detail than that maintained by any single agency or available from any single survey. The 1973 Current Population Survey-Administrative Record Exact Match Study did exactly that, creating an extremely rich data file. Survey records for people in the March 1973 Current Population Survey were linked to their earnings and benefit information in Social Security Administration records as well as to data from their 1972 income tax returns. The file contains the usual Current Population Survey demographic and labor force items plus the March supplement questions on income and work experience, longitudinal data on earnings from the individual's summary earnings record at the Social Security Administration, data on taxable income of tax units from the Internal Revenue Service, and beneficiary status from the Social Security Administration.

Major methodological difficulties had to be solved in implementing this procedure because the various files are maintained for different units. The Social Security

---

[15] In 1977 the Federal Committee on Statistical Methodology formed a subcommittee on statistical uses of administrative records. The report of that subcommittee, now in draft form, looks at achievements, prospects, and problems in the use of administrative data for statistical purposes (see Federal Committee on Statistical Methodology, 1980b).

Administration uses the individual covered worker; the Current Population Survey uses the household, with information solicited for each individual; and the Internal Revenue Service uses the taxpayer, who may be an individual, a couple, or a family. These files are available for public use, and many substantive and methodological studies have used them (see Studies from Interagency Data Linkages, 1980). Early studies during 1975-1978 dealt with methodological issues and cross-sectional analyses of income data. More recent studies are using the files to carry on mortality and disability research (see DelBene and Scheuren, 1979). Little attention has been paid to date to the longitudinal aspects of the files, however.

Two linkage projects on a much larger scale are now in the planning stage. The first, the Linked Administrative Statistical Sample project starts with the Continuous Work History Sample that has been maintained by the Social Security Administration for over 40 years. The Continuous Work History Sample is a 1 percent sample of Social Security accounts, updated each year so that the file is a longitudinal one. (It has been made available to outside researchers; the Tax Reform Act of 1976 has foreclosed the release of information more recent than that date to the public, but the file continues to be updated.) Over the years the file has been used for internal research at the Social Security Administration to keep track of the characteristics of workers covered by Social Security and how this population has changed with legislative changes; it has been used outside the Social Security Administration for research on work force characteristics, life-cycle earnings, and industrial and environmental health issues (Kilss et al., 1980). The purpose of the Linked Administrative Statistical Sample is to supplement the longitudinal data of the Continuous Work History Sample on earnings and benefits histories with mortality information from the National Center for Health Statistics and individual income tax items obtained from the Internal Revenue Service Statistics of Income Program. The long-term goals of this effort are to develop a source of socioeconomic and job-related mortality and morbidity data that might eventually make it possible to separate the residence and occupational influences on health; to construct baseline data on income for small areas to measure the impact of changes in tax policy; and to study regional labor market conditions.

The Survey of Income and Program Participation, a longitudinal survey that went through experimental stages but is currently unfunded, uses administrative records as frames for sampling recipients of programs of income supplementation, such as Aid to Families with Dependent Children or Supplementary Social Security benefits. These administrative records will also be matched with the survey results to enrich the data and will be used as a control to study the accuracy of income reporting by those surveyed, thus adding to knowledge of response errors. The eventual aim of the survey is to support policy analysis of a wide range of federal and state transfer and service programs. Public use data tapes are envisioned (see Griffith and Kasprzyk, 1980).

Other Aspects of the Use of Administrative Data

There are several problems in the use of administrative data for statistical purposes. One is that data collected by an administrative agency as a by-product of its necessary data collection activities are probably less accurate than the primary data. Agency priorities usually involve carefully refining and monitoring of the data that are necessary for programs of that agency, while fewer resources are invested in caring for pieces of information that are less important to the agency itself. Thus, for example, some 11 percent of the workers in the 1 percent Continuous Work History Sample were found to be miscoded on location of work (Cartwright, 1978). For purposes of record keeping in the Social Security Administration, what matters is the employer for whom the employee works, not the location of the work. This presents no problem if an employer has only one place of business, but problems arise with a multiestablishment employer and a system of optional reporting of place of employment. Problems of a similar type will exist whenever files from different agencies are linked, as long as definitions of variables are not the same across agencies. Advances will occur when definitions are standardized.

Legal and ethical problems also exist in the use of administrative records for statistical purposes. Legal restraints and confidentiality rulings differ across agencies. With certain exceptions, the 1976 Tax Reform Act makes it illegal for the Internal Revenue Service to release data except for purposes of tax administration. Other agencies are governed by comparable legislative requirements for confidentiality; in the development of the Linked Administrative Statistical Sample, inconsistent regulations are reconciled by adhering to the more stringent of the two.

The principle of the functional separation of data for statistical and administrative uses has been proposed to deal with confidentiality problems (L. Alexander, 1979). Under such a principle, data that are to be used for statistical purposes may flow from administrative agencies, but there is to be no reverse flow. It is also useful to recognize the difference between human beings and other legal entities in their needs for confidentiality (Federal Committee on Statistical Methodology, 1980b). A continuing problem is that files created for statistical purposes offer the nightmarish possibility of corruption; solutions may lie with the insulated link file system discussed earlier under longitudinal surveys.

One other objection to the use of administrative records for some statistical purposes is that when they are used for program evaluation they may be distorted in order to present a more favorable view of the program (Campbell, 1979). Thus, for example, when the success of a job placement program is evaluated by the number of placements made, there is a tendency to concentrate placement efforts on the clients who are easiest to place. Similarly, when the performance of a police department is measured by percentage of cases cleared, crimes may not be recorded when they are reported but only when and if they are cleared.

An exciting development related to the use of administrative records grows out of the literature on imputation for missing data. The linkages between data sets discussed above operate, through the use of such identifiers as social security numbers, to link data related to the same individual or family from the files of two separate administrative agencies. This is called exact matching. There is also a concept of statistical matching (Radner, 1978), wherein a single data set is created from two sets that do not refer to the same people, in order to create a more comprehensive or accurate set of variables. This is usually done between a household sample survey and an administrative (e.g., tax return) sample or between two surveys. Matching variables are chosen, in the same sorts of ways as adjustment classes are constructed for imputing missing values, and the closest match, in some sense, for each member of one file is chosen from the members of the other file. The files are then merged and the matched data can be treated as if they were measured for individuals for some analytic purposes.

The worth of such a synthetic data file rests on the similarity of definition and error structure of the matching variables across the original files. Statistical matching is a potentially rich and useful source of detailed data, but the accuracy of estimates made from such files remains to be explored. One proposed method for this exploration, reminiscent of methods proposed to evaluate imputation methods, starts with making estimates from a complete data file that is then artificially broken apart. Various methods of statistical matching could be applied to the fragmented file, and the estimates derived from the statistically matched files could be compared with each other and with those derived from the original complete file (see Federal Committee on Statistical Methodology, 1980a).

AND THE FUTURE?

Having examined some of the recent advances in methods for large-scale surveys and experiments, we are perhaps in a position to see what gave rise to such innovations and thus to anticipate what will stimulate the innovations of the next several years, if not what such innovations will be.

The availability of data has a tendency to generate new techniques suitable for their analysis. We have seen how publicly available longitudinal data files have already stimulated the application of mathematical models to increase our understanding of processes. As more longitudinal data are accumulated, new statistical and computational methods will be developed for their organization. As they are organized longitudinally and made available to researchers, we can expect to see further advances in models and methods for their analysis.

Similarly, we now have large archives of survey data collected using multistage stratified cluster sampling techniques, but most of our methods for analyzing these data must squeeze them into a mold designed for less complicated stratified samples. There is a need to develop analytic methods tailored to these complicated sampling designs and a parallel need to experiment with sampling designs that lend themselves to analysis using available multivariate methods.

Available data coupled with policy-relevant results from their analysis seem to constitute a particularly potent force in stimulating creative reanalysis, scientific controversy, and methodological progress. Reanalyses of the evaluation of Head Start programs inspired criticism of some widely used techniques of statistical adjustment and gave impetus to the development of new quasi-experimental designs for the accomplishment and analysis of such evaluations (Boruch, 1980; Cook and Campbell, 1979). As results of some of the current large-scale studies accumulate, they will inevitably create controversy and innovation.

We have seen that methodological techniques have transferability. Experimental design was developed in agriculture, adapted to laboratory experiments in the natural and behavioral sciences, and adapted again to the field situations that are social experiments. But with each adaptation new problems were faced; their solutions constitute some of the methodological advances chronicled above. If we can envision new fields for experimentation, we can perhaps anticipate methodological developments. Energy consumption, conservation, and conversion come readily to mind as focal areas for policy discussion. Although few experiments have been mounted in these areas (the Los Angeles Peak Load Electricity Pricing Experiment is a notable exception), they seem ripe for research seeking to inform policy. Moreover, there is movement toward involving social scientists other than economists in understanding these aspects of energy (for example, a committee formed by the National Research Council on the behavioral and social aspects of energy consumption). This combination of a relatively unexplored field and some novel points of view may pose new problems and result in methodological innovation.

The very policy relevance of the issues addressed by large-scale surveys and experiments has inspired and should continue to inspire innovation through the communication of results. Complicated methodological techniques may be discussed elliptically between involved professionals, sometimes with little examination of the underlying assumptions and their implications for conclusions. When, however, the results of policy-relevant research must be discussed with policy makers and the general public and the question "How do you know?" responded to, careful explanation in understandable terms becomes crucial. Such explanation can lead to better understanding of the strengths and weaknesses of techniques and to their revision. Similarly, attempts to communicate results have been contributing factors to the recent renaissance in statistical graphics (e.g., Fienberg, 1979). This renaissance, in turn, has inspired programs of developing and evaluating the effectiveness of innovative graphic displays (e.g., Kruskal, 1980; Wainer and Francolini, 1980).

The process is a disorderly cumulation; to meet a perceived need for information or to cope with available data we use or adapt methods from the scientific warehouse. This attempt to transfer methods and/or the discovery of their flaws leads to improved methods and the generation of new data. These are then available in the scientific warehouse both to stimulate and to be used in the next cycle of information seeking.

ACKNOWLEDGMENTS

Thanks are due to a great many people for enormous amounts of assistance during the preparation of this paper. First of all to the advisory committee, who commented inspiringly through interminable iterations of outlines and drafts: William H. Kruskal, Stephen E. Fienberg, Norman M. Bradburn, and Richard A. Berk. Officers and staff at the Social Science Research Council were also unstinting in time devoted to intellectual guidance and moral support: David L. Sills, Peter B. Read, Kenneth Prewitt, and Roberta B. Miller. In trying to describe the cutting edge of research, I had to depend in large part on the willingness of investigators to share their work in progress, and a great many people responded generously, too many to list here. The blanket nature of such thanks should not be read as a detraction from its warmth. Many friends and colleagues with no official connection to the project gave generously of their time to explain and instruct in their areas of special competence and to make comments on various drafts (several of these people might well be considered honorary members of the advisory committee: Eugene Weinstein, Katherine Wallman, Charles Turner, Wendell Thompson, Edward Tufte, Frank Stafford, Charles Smith, Burton Singer, Merrill Shanks, Karl Schuessler, Frederick Scheuren, Gordon Sande, Ingram Olkin, Frederick Mosteller, Susan Miskura, Albert Madansky, Richard Link, Michael Kagay, Edwin Goldfield, Jonathan Cole, Robert Boruch, Albert Biderman, and Barbara Bailar. The errors and infelicities, of course, are my very own. Linda D. Anderson performed miracles of typing, both in speed and in beauty.

REFERENCES

Aigner, Dennis J.
1979 "A brief introduction to the methodology of optimal experimental design." Journal of Econometrics 11:7-26.

Alexander, C. Norman, and G. W. Knight
1971 "Situated identities and social psychological experimentation." Sociometry 34:65-82.

Alexander, Lois
1979 "Statistical progeny of administrative records: some legal issues." In Linda DelBene and Fritz Scheuren, eds., Statistical Uses of Administrative Records With Emphasis on Mortality and Disability Research. Washington, D.C.: U.S. Department of Health, Education, and Welfare, Social Security Administration.

Andersen, Ronald, Judith Kasper, Martin R. Frankel, and associates
1979 Total Survey Error. San Francisco: Jossey-Bass.

Anderson, Sharon, Ariane Auquier, Walter W. Huack, David Oakes, Walter Vandaele, and Herbert I. Weisberg
1980 Statistical Methods for Comparative Studies. New York: Wiley.

Archibald, Rae W., and Joseph P. Newhouse
1980 Social Experimentation: Some Whys and Hows. R-2479-HEW. Santa Monica, Calif.: Rand Corp.

Aziz, Faye, and Fritz Scheuren
1978 Imputation and Editing of Faulty or Missing Survey Data. Washington, D.C.: U.S. Department of Commerce, Bureau of the Census. (Papers presented at the 1978 meeting of the American Statistical Association and also printed in the Proceedings volumes for that meeting.)

Bailar, Barbara A.
1976 "Some sources of error and their effects on census statistics." Demography 13:273-286.

Bailar, Barbara A., and John C. Bailar, III
in press "Comparison of the biases of the 'hot deck' imputation procedure with an 'equal-weights' imputation procedure." In Panel on Incomplete Data, Symposium on Incomplete Data. New York: Academic Press.

Bailar, Barbara A., and Susan Miskura
1980 The 1980 Experimental Program. Unpublished paper, U.S. Bureau of the Census.

Bailar, John C., III
1978 Discussion. Pp. 62-64 in Aziz and Scheuren, eds., Imputation and Editing of Faulty or Missing Survey Data. Washington, D.C.: U.S. Department of Commerce, Bureau of the Census.

Bailar, John C., III, and Barbara A. Bailar
1978 "Comparison of two procedures for imputing missing survey values." Pp. 65-75 in Aziz and Scheuren, eds., Imputation and Editing of Faulty or Missing Survey Data. Washington, D.C.: U.S. Department of Commerce, Bureau of the Census.

Berk, R. A., and S. F. Berk
1979 Labor and Leisure at Home. Beverly Hills, Calif.: Sage.

Bernstein, Ilene Nagel, and Howard E. Freeman
1975 Academic and Entrepreneurial Research. New York: Russell Sage Foundation.

Beveridge, Andrew A., and John B. Taylor
1980 Quarterly Report--3rd Quarter on Grant HUD-5516 RG. A Continuation of the Longitudinal Transformation and Analysis of the Annual Housing Survey. New York: Center for the Social Sciences, Columbia University.

Biderman, Albert D., Louise A. Johnson, Jennie McIntyre, and Adrianne W. Weir
1967 Report on a Pilot Study in the District of Columbia on Victimization and Attitudes Toward Law Enforcement. President's Commission on Law Enforcement and Administration of Justice Field Survey--# 1. Washington, D.C.: U.S. Government Printing Office.

Bielby, William T., Clifford C. Hawley, and David Bills
1977 "Research uses of the National Longitudinal Surveys." In A Research Agenda for the National Longitudinal Surveys of Labor Market Experience: Report on the Social Science Research Council's Conference on the National Longitudinal Surveys, October 1977. Washington, D.C.: Social Science Research Council.

Bishop, Yvonne M. M., Stephen E. Fienberg, and Paul W. Holland
1975 Discrete Multivariate Analysis. Cambridge, Mass.: MIT Press.

Boruch, Robert F.
1975 "Contentions about randomized field experiments. In Robert F. Boruch and Henry W. Riecken, eds.,

Experimental Testing of Public Policy: The Proceedings of the 1974 Social Science Research Council Conference on Social Experiments. Boulder, Colo.: Westview Press.

1980 Working Paper; Problems in Social Program Evaluation: Analogs from Engineering and the Physical Sciences, Medicine, and an Assortment of Other Disciplines, Together with Historical Reference. Psychology Department, Division of Methodology and Evaluation Research, Northwestern University.

Boruch, Robert F., and Joe S. Cecil

1979 Assuring the Confidentiality of Social Research Data. Philadelphia: University of Pennsylvania Press.

Box, George E. P.

1978 "Experimental design: response surface methodology." In William H. Kruskal and Judith M. Tanur, eds., International Encyclopedia of Statistics. New York: Free Press.

Bradburn, Norman M.

1978 "Response effects." In Peter H. Rossi, and James D. Wright, eds., The Handbook of Survey Research. New York: Academic Press.

Bradburn, Norman M., Seymour Sudman, and associates

1979 Improving Interview Method and Questionnaire Design: Response Effects to Threatening Questions in Survey Research. San Francisco: Jossey-Bass.

Brewer, Kenneth R., and Carol E. Sarndal

in press "Six approaches to enumerative survey sampling." In Panel on Incomplete Data Symposium on Incomplete Data. New York: Academic Press.

Brooks, Camilla A., and Barbara A. Bailar

1978 An Error Profile: Employment as Measured by the Current Population Survey. Statistical Policy Working Paper 3. Washington, D.C.: U.S. Department of Commerce.

Campbell, Donald T.

1978 "Experimental design: quasi experimental design." In William H. Kruskal and Judith M. Tanur, eds., International Encyclopedia of Statistics. New York: Free Press.

1979 "Assessing the impact of planned social change." Evaluation and Program Planning, 2:67-90.

Campbell, Donald T., and Julian C. Stanley
1963 Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally.

Cannell, C. F.
1977 A Summary of Studies of Interviewing Methodology. Vital and Health Statistics, Series 2, No. 69. Washington, D.C.: U.S. Government Printing Office.

Cannell, Charles F., Lois Oksenberg, and Jean M. Converse, eds.
1978 Experiments in Interviewing Techniques: Field Experiments in Health Reporting, 1971-1977. National Center for Health Services Research, Research Proceedings Series, DHEW Publication No. (HRA) 78-3204. Washington, D.C.: U.S. Department of Health, Education, and Welfare.

Cartwright, David W.
1978 Major Limitations on CWHS Files and Prospects for Improvement. Paper presented at NBER Workshop on Policy Analysis with Social Security Research Files, March 17.

Cochran, William G.
1978 "Experimental design: the design of experiments." In William H. Kruskal and Judith M. Tanur, eds., International Encyclopedia of Statistics. New York: Free Press.

Conlisk, John, and Harold Watts
1969 "A model for optimizing experimental designs for estimating response surfaces." In Proceedings of the American Statistical Association, Social Statistics Section. Reprinted in 1979 in Journal of Econometrics 11:27-42.

Cook, Thomas D., and Donald T. Campbell
1979 Quasi Experimentation: Design and Analysis for Field Settings. Chicago: Rand McNally.

Cox, Brenda G., and Ralph E. Folsom
1978 "An empirical investigation of alternative item nonresponse adjustments." Pp. 51-55 in Aziz and Scheuren, eds., Imputation and Editing of Faulty or Missing Survey Data. Washington, D.C.: U.S. Department of Commerce, Bureau of the Census.

Cox, D.R.
1972 "Regression models and life tables." Journal of the Royal Statistical Society, B. 34:187-220.

Csikszentmihalyi, M., R. Larsen, R., and S. Prescott

1977 "The ecology of adolescent activities and experiences." J. Youth and Adolescence, 6: 281-294.

Dalenius, Tore

1977 "Bibliography of nonsampling errors in surveys." International Statistical Institute Review 45:71-90 (April, A - G); 181-197 (August, H - Q); 303-317 (December, R - Z).

in press "Informed consent or R.S.V.P." In Panel on Incomplete Data, Symposium on Incomplete Data. New York: Academic Press.

Davis, James A.

1975 "Are surveys any good, and if so, for what?" In H. Wallace Sinaiko, and Laurie A. Broedling, eds., Perspectives on Attitude Assessment: Surveys and Their Alternatives. Proceedings of a Conference held at The Bishop's Lodge, Santa Fe, New Mexico, April 22-24. Washington, D.C.: Manpower Research and Advisory Services, Smithsonian Institution.

Deighton, Richard E., James R. Poland, Joel R. Stubbs, and Robert D. Tortora

1978 Glossary of Nonsampling Error Terms. Prepared for the Executive Office of the President, Office of Management and Budget, Federal Committee on Statistical Methodology, Subcommittee on Nonsampling Errors.

DelBene, Linda, and Fritz Scheuren, eds.

1979 Statistical Uses of Administrative Records With Emphasis on Mortality and Disability Research. Washington, D.C.: U.S. Department of Health, Education, and Welfare, Social Security Administration.

Deming, W. Edwards

1978 "Sample surveys: the field." In William H. Kruskal and Judith M. Tanur, eds., The International Encyclopedia of Statistics. New York: Free Press.

Deming, W. Edwards, and Frederick F. Stephan

1940 On a least squares adjustment of a sampled frequency table when the expected marginal tables are known. Annals of Mathematical Statistics 11:427-444.

Deutscher, Irwin

1973 What We Say/What We Do. Glenview, Ill.: Scott, Foresman.

Duncan, Otis Dudley, and Howard Schuman
1980 Effects of question wording and context: an experiment with religious indicators. Journal of the American Statistical Association 75: 269-275.

Durako, Stephen, and Thomas McKenna
1980 Collecting Health Interview Survey Data by Telephone: A Mailout Experiment. Paper prepared for the annual meeting of the American Statistical Association, August 11-14.

Dutka, Solomon, and Lester R. Frankel
1980 Sequential Survey Design Through the Use of Computer Assisted Telephone Interviewing. Paper presented at the annual meeting of the American Statistical Association, August 11-14.

Elandt-Johnson, Regina C., and Norman L. Johnson
1980 Survival Models and Data Analysis. New York: Wiley.

Ennis, Philip H.
1967 Criminal Victimization in the United States: A Report of a National Survey. Chicago: National Opinion Research Center.

Federal Committee on Statistical Methodology, Subcommittee on Matching Techniques
1980a Report on Exact and Statistical Matching Techniques. Statistical Policy Working Paper 5. Washington, D.C.: U.S. Department of Commerce, Office of Federal Statistical Policy and Standards.

Federal Committee on Statistical Methodology, Subcommittee on Statistical Uses of Administrative Records
1980b Report on Statistical Uses of Administrative Records. Washington, D.C.: U.S. Department of Commerce, Office of Federal Statistical Policy and Standards.

Ferber, Robert, and Werner Z. Hirsch
1979 "Social experiments in economics." Journal of Econometrics 11:77-115.

Ferber, Robert, Paul Sheatsley, Anthony Turner, and Joseph Waksberg
1980 What is a Survey? Subcomcommittee of the Section on Survey Research Methods, Washington D.C.: American Statistical Association.

Fienberg, Stephen E.
1979 "Graphical methods in statistics." The American Statistician 33:165-178.

Fienberg, Stephen E.
1980a The Measurement of Crime Victimization: Prospects for Panel Analysis of a Panel Survey. Paper presented at "Censuses and Sample Surveys," Institute of Statisticians International Conference, Trinity College, Cambridge, England, July 2-5.
1980b "Statistical modelling in the analysis of repeat victimization." In Stephen E. Fienberg and Albert J. Reiss, Jr., eds., Indicators of Crime and Criminal Justice: Quantitative Studies. Washington, D.C.: U.S. Government Printing Office.
Fienberg, Stephen E., Kinley Larntz, and Albert J. Reiss, Jr.
1976 "Redesigning the Kansas City Preventive Patrol Experiment." Evaluation 3:124-131.
Ford, Barry L.
1976 Missing Data Procedures: A Comparative Study. U.S. Department of Agriculture, Sampling Studies Section, Sample Surveys Research Branch, Statistical Reporting Service, Washington D.C.
1978 Missing Data Procedures: A Comparative Study. Part 2. U.S. Department of Agriculture, Sampling Studies Section, Sample Surveys Research Branch, Statistical Research Division.
Franklin, Paula
1979 "Planning for program experimentation for the Social Security Disability Insurance Program." In Proceedings of the American Statistical Association, Social Statistics Section. Also in Linda DelBene and Fritz Scheuren, eds., Statistical Uses of Administrative Records with Emphasis on Mortality and Disability Research. Washington, D.C.: U.S. Department of Health, Education, and Welfare, Social Security Administration.
Freedman, Deborah S., Arland Thornton, and Donald Camburn
1980 "Maintaining response rates in longitudinal Studies." Sociological Methods and Research 9:87-98.
Freeman, Howard E.
1980 Research Opportunities Related to CATI. Paper prepared for the University of California Conference on Computer Assisted Survey Technology.

Gilbert, John P., Richard J. Light, and Frederick Mosteller

1975 "Assessing social innovations: an empirical base for policy." In A. Lumsdaine and C. A. Bennett, eds., Central Issues in Social Program Evaluation. New York: Academic Press.

Granger, R. L., et al.

1969 The Impact of Head Start, An Evaluation of the Effects of Head Start on Children's Cognitive and Affective Development. Vol. 1. Report to the U.S. Office of Economic Opportunity by Westinghouse Learning Corporation and Ohio University.

Griffith, Jeanne E., and Daniel Kasprzyk

1980 "The use of administrative records in the Survey of Income and Program Participation." In Federal Committee on Statistical Methodology, Subcommittee on Statistical Uses of Administrative Records, Report on Statistical Uses of Administrative Records. Washington, D.C.: U.S. Department of Commerce, Office of Federal Statistical Policy and Standards.

Griliches, Z., B. H. Hall, and J. A. Hausman

1977 Missing Data and Self Selection in Large Panels. Paper presented at the INSEE Conference "Economics of Panel Data," Paris, August.

Groves, Robert M., and Robert L. Kahn

1979 Surveys by Telephone. New York: Academic Press.

Groves, Robert M., and Lou J. Magilavy

1980 Effects of Interviewer Variance in Telephone Surveys. Paper prepared for presentation at the annual meeting of the American Statistical Association, August 11-14.

Groves, Robert M., Marianne Berry, and Nancy Mathiowetz

1980 Some Impacts of Computer Assisted Telephone Interviewing on Survey Methods. Paper prepared for presentation at the annual meeting of the American Statistical Association, August 11-14.

Hausman, Jerry A., and David Wise

1977 "Social experimentation, truncated distributions, efficient estimation." Econometrica 45:919-938.

Heckman, James D.

1976 "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models." Annals of Economic and Social Measurement 5:475-492.

1979 "Sample selection bias as specification error." Econometrica 47:153-161.

Herzberg, Agnes, and D. R. Cox

1969 "Recent work on the design of experiments: a bibliography and a Review." Journal of the Royal Statistical Society 132:29-67.

Hill, Martha S., and F. Thomas Juster

1979 Constraints and Complementarities in Time Use. Discussion draft. Survey Research Center, University of Michigan.

Hill, W. J., and W. G. Hunter

1966 A review of response surface methodology: a literature survey. Technometrics 8:571-590.

Hoff, N. Gail, and Marvin M. Thompson

1980 Diaries in a Consumer Expenditure Survey. Paper presented at the annual meeting of the American Statistical Association, August 11-14.

Horvitz, Daniel G.

1980 On the Significance of a Survey Design Information System. Paper presented at the annual meeting of the American Statistical Association, August 11-14.

Jöreskog, Karl G., and Sörbom, Dag

1979 Advances in Factor Analysis and Structural Equation Models. Jay Magidson, ed. Cambridge, Mass.: Abt Books.

Kahn, Robert L., and Charles F. Cannell

1978 "Interviewing in social research." In William H. Kruskal and Judith M. Tanur, eds., International Encyclopedia of Statistics. New York: Free Press.

Kalachek, Edward

1979 "Longitudinal surveys and labor market analysis." In Data Collection, Processing, and Presentation, Vol. II of Appendix of Counting the Labor Force, Report of the National Commission on Employment and Unemployment Statistics. Washington, D.C.: U.S. Government Printing Office.

Kalbfleisch, John D. and Ross L. Prentice

1980 The Statistical Analysis of Failure Time Data. New York: Wiley.

Kalsbeek, William D.

1980 A Conceptual Review of Survey Error Due to Nonresponse. Paper presented at the annual meeting of the American Statistical Association, August 11-14.

Kalton, Graham, and Howard Schuman
1980 The Effect of the Question on Survey Response. Paper presented at the annual meeting of the American Statistical Association, August 11-14.

Kilss, Beth, Fritz Scheuren, and Warren Buckler
1980 Goals and Plans for a Linked Administrative Statistical Sample. Paper presented at the annual meeting of the American Statistical Association, August 11-14.

Kish, Leslie
1965 Survey Sampling. New York: Wiley.

Kohn, Melvin L., and Carmi Schooler
1978 "The reciprocal effects of the substantive complexity of work and intellectual flexibility: a longitudinal assessment." American Journal of Sociology 84:24-52.

Kruskal, William H.
1980 Criteria for Statistical Graphics. Paper presented at the annual meeting of the American Statistical Association, August 11-14.

Lebby, D. Edwin
1980 CATI's First Decade: The Chilton Experience. Paper prepared for the University of California Conference on Computer Assisted Survey Technology.

Lessler, Judith R.
in press "An expanded survey error model." In Panel on Incomplete Data, Symposium on Incomplete Data. New York: Academic Press.

Little, Roderick J. A., and Donald B. Rubin
in press "Six approaches to enumerate survey sampling. Discussion." In Panel on Incomplete Data, Symposium on Incomplete Data. New York: Academic Press.

Locander, William, Seymour Sudman, and Norman Bradburn
1974 "An investigation of interview method, threat and response distortion." *Proceedings of the American Statistical Association, Social Statistics Section*:21-27.

Madansky, Albert
1980 Response Surface Exploration in Social Experiments. Paper presented at the Second Annual Public Policy Conference, Boston, Mass., October 24-25.

Morgan, James N.
1977 Individual Behavior, Economic Analysis, and Public Policy. The 1977 Wladmir Woytinsky Lecture.

Morris, Carl
1975 "A finite selection model for experimental design of the Health Insurance Study." Proceedings of the American Statistical Association, Social Statistics Section. Reprinted in 1979 in Journal of Econometrics 11:43-61.
in press "Nonresponse issues in public policy experiments, with emphasis on the Health Insurance Study." In Panel on Incomplete Data, Symposium on Incomplete Data. New York: Academic Press.

Moses, Lincoln E.
1978 "Statistical analysis, special problems of: truncation and censorship." In William H. Kruskal and Judith M. Tanur, eds., International Encyclopedia of Statistics. New York: Free Press.

Mosteller, Frederick
1978 "Nonsampling errors." In William H. Kruskal and Judith M. Tanur, eds., International Encyclopedia of Statistics. New York: Free Press.

Mosteller, Frederick, and Daniel P. Moynihan, eds.
1972 On Equality of Educational Opportunity. Papers derived from the Harvard University Faculty Seminar on the Coleman Report. New York: Vintage Press.

Neter, J., and J. Waksberg
1964 "A study of response errors in expenditures data from household interviews." Journal of the American Statistical Association 59:18-55.

Nicholls, William L. II, George A. Lavender, and J. Merrill Shanks
1980 An Overview of Berkeley SRC CATI, Version 1. Survey Research Center Working Paper 31, University of California, Berkeley.

Nicholson, W., and S. R. Wright
1977 "Participants' understanding of treatment in policy experimentation." Evaluation Quarterly 1:171-186.

Nisselson, Harold, and Theodore D. Woolsey
1959 "Some problems of the household interview design for the National Health Survey." Journal

of the American Statistical Association 54:69-87.

Oh, H. Lock, and Fritz Scheuren
1978 "Multivariate raking ratio estimation in the 1973 Exact Match Study and some unresolved application issues in raking ratio estimation." Pp. 120-135 in Aziz and Scheuren, eds., Imputation and Editing of Faulty or Missing Survey Data Washington, D.C.: U.S. Department of Commerce, Bureau of the Census.

Orne, Martin T.
1962 "On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implication. American Psychologist 17:776-783.

Panel on Incomplete Data of the Committee on National Statistics, National Research Council.
in press Symposium on Incomplete Data. New York: Academic Press.

Panel on Privacy and Confidentiality as Factors in Survey Response, Committee on National Statistics, National Research Council
1979 Privacy and Confidentiality as Factors in Survey Response. Washington, D.C.: National Academy of Sciences.

Pullum, Thomas
1978 "Postscript to social mobility." In William H. Kruskal and Judith M. Tanur, eds., International Encyclopedia of Statistics. New York: Free Press.

Quay, Herbert C.
1977 "The three faces of evaluation: what can be expected to work." Criminal Justice and Behavior 4:341-354. Reprinted in Lee Sechrest et al., eds. Evaluation Studies Review Annual. Vol. 4. Beverly Hills, Calif.: Sage.

Radner, Daniel B.
1978 "The development of statistical matching in economics." In Aziz and Scheuren, eds., Imputation and Editing of Faulty or Missing Data. Washington, D.C.: U.S. Department of Commerce, Bureau of the Census.

Ramsoy, Natalie Rogoff, and Sten-Erick Clausen
1978 "Events as units of analysis in life history studies." In A Research Agenda for the National Longitudinal Surveys of Labor Market Experience: Report of the Social Science Research Council

on the National Longitudinal Surveys. October 1977.

Ray, Subhash C., Richard A. Berk, and William T. Bielby
1980 Correcting Sample Selection Bias for Bivariate Logistic Distribution of Disturbances. Working Paper in Economics #160, University of California, Santa Barbara, Department of Economics.

Reichhardt, Charles S.
1979 The Statistical Analysis of Data from Nonequivalent Group Designs. In Thomas D. Cook and Donald T. Campbell, eds., Quasi-Experimentation. Chicago: Rand McNally.

Reiss, Albert J., Jr.
1980 "Victim proneness by type of crime in repeat victimization." In Stephen E. Fienberg and Albert J. Reiss, Jr., eds., Indicators of Crime and Criminal Justice: Quantitative Studies. Washington, D.C.: U.S. Government Printing Office.

Riecken, Henry W., and Robert F. Boruch, eds.
1974 Social Experimentation: A Method for Planning and Evaluating Social Intervention. New York: Academic Press.

Rosenthal, Robert
1966 Experimenter Effects in Behavioral Research. New York: Appleton-Century-Crofts.

Roshwalb, Irving, Leonard Spector, and Albert Madansky
1979 New Methods of Telephone Interviewing A&S/CATI. Proceedings of the XXXII Esomar Congress, The Challenge of the Eighties. Brussels, Belgium, September 2-6.

Rubin, Donald B.
1978 "Multiple imputations in sample surveys--a phenomenological Bayesian approach to nonresponse. Followed by a discussion and rejoinder." Pp. 1-18 in Aziz and Scheuren, eds., Imputation and Editing of Faulty or Missing Data. Washington, D.C.: U.S. Department of Commerce, Bureau of the Census.
1979 Handling Nonresponse in Sample Surveys by Multiple Imputations. Monograph prepared for the Census Bureau.

Sande, Innis G.
in press (a) "Hot deck imputation procedures." In Panel on Incomplete Data, Symposium on Incomplete Data. New York: Academic Press.

Sande, Gordon
in press (b) "Hot deck discussion-replacement for a ten minute gap." In Panel on Incomplete Data, Symposium on Incomplete Data. New York: Academic Press.

Schuman, Howard, and Stanley Presser
1978 "The assessment of 'no opinion' in attitude surveys." In Karl F. Schuessler, ed., Sociological Methodology 1979. San Francisco: Jossey-Bass.

Sechrest, Lee, and Robin Redner
1979 "Strength and integrity of treatments in evaluation studies." In How Well Does It Work: Review of Criminal Justice Evaluation 1978. National Criminal Justice Reference Service, National Institute of Law Enforcement and Criminal Justice.

Sechrest, Lee, Stephen G. West, Melinda A. Phillips, Robin Redner, and William Yeaton
1979 "Introduction." In Lee Sechrest et. al., eds., Evaluation Studies Review Annual. Volume 4. Beverly Hills: Sage.

Shanks, J. Merril
1980 The Development of CATI Methodology. Paper prepared for the University of California Conference on Computer Assisted Survey Technology.

Shimizu, I.M., and G. S. Bonham
1978 "Randomized response technique in a national survey." Journal of the American Statistical Association 73:35-39.

Simon, Herbert A.
1980 The Behavioral and Social Sciences. Science 209:72-78.

Singer, Burton, and Seymour Spilerman
1976 "Some methodological issues in the analysis of longitudinal surveys." Annals of Economic and Social Measurement 5: 447-474.
1977 "Fitting stochastic models to longitudinal survey data--some examples in the social sciences." Bulletin of the International Statistical Institute 47:283-300.

Singer, Eleanor
1978 "Informed consent: consequences for response rate and response quality in social surveys." American Sociological Review 43:144-162.

Sirkin, Monroe G.
1975 Alcohol and Other Drug Use and Abuse in the State of Michigan. Office of Substance Abuse Services, Michigan Department of Public Health, April.

Stafford, Frank, and Greg J. Duncan
1979 The Use of Time and Technology by Households in the United States. Working Paper of the Institute for Social Research, University of Michigan.

Stephenson, C. B.
1978 A Comparison of Full-Probability and Probability--with Quotas Sampling Techniques in the General Social Survey. GSS Technical Report No. 5, National Opinion Research Center, Chicago. (Forthcoming in Public Opinion Quarterly.)
1980 Studies from Interagency Data Linkages, Report No. 11: Measuring the Impact on Family and Personal Income Statistics of Reporting Differences between the Current Population Survey and Administrative Sources. SSA Publ. No. 13-11750. U.S. Department of Health, Education, and Welfare, Social Security Administration.

Sudman, Seymour
1967 Reducing the Cost of Surveys. Chicago: Aldine.

Sudman, Seymour, and Norman N. Bradburn
1973 "Effects of time and memory factors on responses in surveys." Journal of the American Statistical Association, 68:805-815.
1974 Response Effects in Surveys: A Review and Synthesis. Chicago: Aldine.

Sudman, Seymour, and Robert Ferber
1971 "Experiments in obtaining consumer expenditure by diary methods." Journal of the American Statistical Association 66:725-735.

Sudman, Seymour, Edward Blair, Norman M. Bradburn, and Carol Stocking
1977 "Estimates of threatening behavior based on reports of friends." Public Opinion Quarterly 41:261-264. Reprinted in Norman M. Bradburn and Seymour Sudman and associates (1979) Improving Interview Method and Questionnaire Design: Response Effects to Threatening Questions in Survey Research. San Francisco: Jossey-Bass.

Tharp, Roland G., and Ronald Gallimore
1979 "The ecology of program research and

evaluation: a model of evaluation succession." In Lee Sechrest et al., eds., Evaluation Studies Review Annual. Volume 4. Beverly Hills, Calif.: Sage.

Thompson, Wendel L., Lynda T. Carlson, Thomas H. Woteki, and Kenneth A. Vagts
1980 Improving the Quality of Data from Monthly Gasoline Purchase Diaries. Paper presented at the annual meeting of the American Statistical Association, August 11-14.

Thornberry, Owen I., Jr., and James T. Massey
1978 "Correcting for undercoverage bias in random digit dialed national health surveys." Pp. 56-61 in Aziz and Scheuren, eds., Imputation and Editing of Faulty or Missing Data. Washington, D.C.: U.S. Department of Commerce, Bureau of the Census.

Tuma, Nancy Brandon, Michael T. Hannan, and Lyle P. Groeneveld
1979 "Dynamic analysis of event histories." American Journal of Sociology 84:820-854.

Tupek, Alan R., and W. Joel Richardson
1978 "Use of ratio estimates to compensate for nonresponse bias in certain economic surveys." Pp. 24-29 in Aziz and Scheuren, eds., Imputation and Editing of Faulty or Missing Data. Washington, D.C.: U.S. Department of Commerce, Bureau of the Census.

Turner, Charles F.
1981 "Surveys of Subjective Phenomena: A Working Paper." Pp. 37-78 in Dennis Johnston, ed., Measurement of Subjective Phenomena. U.S. Special Demographic Analyses CDS-803. Washington, D.C.: U.S. Department of Commerce, Bureau of the Census.

Turner, Charles F., and Elissa Krauss
1978 "Fallible indicators of the state of the nation." American Psychologist 33:456-470.

Turner, Charles F., and Elizabeth Martin, eds.
1981 Surveys of Subject Phenomena: Summary Report. Panel on Survey Measurement of Subjective Phenomena, Committee on National Statistics, National Research Council. Washington, D.C.: National Academy Press.
forthcoming Surveying Subjective Phenomena. 2 volumes.

U.S. National Health Survey

1963 Comparison of Hospitalization Reporting in Three Survey Procedures, A Study of Survey Methods for Collection of Hospitalization Data from Household Respondents. Washington, D.C.: U.S. Department of Health, Education, and Welfare, Public Health Service. By Charles Cannell and Floyd Fowler and republished in Vital and Health Statistics, Ser. 2, No. 8, July 1965.

Wainer, Howard, and Carl M. Francolini

1980 "An empirical inquiry concerning human understanding of two-variable color maps." The American Statistician 34:81-93.

Wallman, Katherine K.

1980 Statistics and the Allocation of Federal Funds. Paper presented to the Federal Statistics Users Conferences, Washington, D.C., November 19.

Warner, Stanley

1965 "Randomized response: survey technique for eliminating evasive answer bias." Journal of the American Statistical Association 60:63-69.

Woltman, Henry F., Anthony G. Turner, and John M. Bushery

1980 "A comparison of three mixed-Mode interviewing procedures in the National Crime Survey." Journal of the American Statistical Association 75:534-543.

Research in Psychophysics

L. D. Braida, Tom N. Cornsweet,
N. I. Durlach, David M. Green, Herschel
Leibowitz, Alvin Liberman, R. Duncan Luce,
Richard Pew, and Carl Sherrick

INTRODUCTION

Psychophysics as a discipline arose during the middle of the last century from the burgeoning success of physics in explaining natural phenomena. It was thought that the methodology used so effectively in dealing with the relationships of material things to each other--at that time being translated into the industrial revolution--could also be used to characterize the relationships between the material and the mental worlds. At the beginning the movement attempted to encompass a wide range of phenomena, from the just detectable difference in weight of objects to the beauty of works of art. The lasting heritage from that era of psychophysics is the acceptance of the scientific procedures of the physical sciences in psychology: fully defined and replicable experimental methods, results expressed in numerical terms, the formulation of theories in the form of mathematical functions, and the use of these theories to extrapolate (i.e., make predictions) to specified new situations.

As it is seen today, the goal of psychophysics is to understand at the most basic level the ways in which one apprehends changes in the physical environment, especially when one is attending to such changes. Often the simplest possible changes in stimulation are used--a shift in the energy level of a spot of light or a change in the frequency of a pure tone. Such simple stimuli are usually referred to as signals. In other studies more complex stimuli are used--the presentation of alternating bands of dark and light on the whole visual field or spoken words. When the stimuli are very simple, we usually speak of visual, auditory, or tactile sensation. As the stimuli become more complex and the questions become more ones of

380

grouping and classifying stimuli rather more than just detecting them or telling them apart, we begin to speak of visual, auditory, or tactile perception. Psychophysics incorporates all of the work on sensation and some of the simpler aspects of perception; however, the boundary line between sensation and perception is rather fuzzy. Reading is a perceptual process that very few would class as psychophysical, but just how one perceives individual letters and even how one groups them into words many construe as part of psychophysics. In sum, then, psychophysics is the study of how one makes distinctions about the energy impinging on one, and it attempts to understand one's abilities to extract information from the environment. When the focus of interest becomes that of extracting the meaning embodied in the stimuli, one is outside the bounds of psychophysics and into the general domain of cognitive psychology.

Is psychophysics just a single name for the study of the several senses? Not really, because there is considerable interest in comparing systems and in finding common mechanisms. Such mechanisms may arise for two reasons. First, nature often repeats its solutions, e.g., the same technique in the peripheral nervous system for enhancing differences appears in several of the senses. Second, much of what is interesting about the way information is extracted appears to be mediated not only by the sensory transducer--ear, eye, taste buds, etc.--but also by the brain itself, and some of the higher processing may use the same brain mechanisms for different senses.

Our discussion is divided into two major parts. In the first we treat basic research in psychophysics and in the second some of its applications. The basic research is separated into four parts: the psychophysics and physiology of the visual system, with major emphasis on the sensory aspect; the same for the auditory system; the common problems that arise when cognitive factors, such as instructions, play a crucial role in how the subject responds to the questions posed; and the study of the motor responses of a person in dynamic interaction with the sensory environment, as in flying. Several of the sensory processes--touch, taste, pain, and heat--are not dealt with explicitly in this first part because considerably less is known about them than about vision and audition. (Some aspects of that work, especially for touch, are described in the section on applications.) The section on applications is concerned with attempts to classify and to overcome serious sensory deficits, except for the first subsection, which is concerned with the problem of why drivers think they can see obstacles better at night than they really can.

Much of the work we are mentioning is inherently rather technical. We attempt to keep jargon to a minimum, but some technical concepts are needed. We often attempt to suggest them by example or illustration rather than offer explicit definitions. For truly precise definitions and statements of results, the reader is urged to consult the literature. One general point of caution. It is essential to maintain a distinction between the physical attributes that are manipulated experimentally and the subjective sensations that are related to these physical variables. For example, as experimenters we can manipulate two features of a pure tone, its amplitude (or intensity) and its frequency. The listener can speak of and respond to questions about its loudness and pitch. Loudness is a sensation that is mainly affected by the signal amplitude, but it is also influenced by frequency, which is why one has a loudness control as well as a volume control on good amplifiers. And pitch is affected mainly by the signal frequency, but it is also affected to a lesser degree by the amplitude. Thus, loudness and intensity are not interchangeable, nor are pitch and frequency. The same is true of luminance and brightness, sugar concentration and sweetness, etc.

BASIC RESEARCH IN PSYCHOPHYSICS

Vision

The optical surfaces of the two eyes interact with light reflected from surrounding objects to form images of those objects on a layer of tissue inside each eye that converts light into electrochemical signals. The images have the same properties as those formed on the film of an ordinary camera, and the chemicals in the eye respond to the light in ways that are closely analogous to the ways in which the corresponding chemicals respond in film, producing what is essentially a chemical picture of the scene. However, within a few thousandths of a second after this picture is formed, the neural elements in the eye begin to transform it through processes that are very different from anything in a camera. Parts of the neural processes are magnified and others are drastically reduced. Separations are made, too, so that different aspects of the image are conducted to and processed by different sets of neural mechanisms. Then the neural images are conducted out of the eye and into various regions of the brain, where further, more complex processing occurs. These neural processes strongly affect what we see in ways that are only now becoming understood.

For example, the fact that a lemon looks yellow and bright whether it is seen inside a kitchen or outside on a tree in broad daylight may not seem surprising. However, if it were not for some very complex processing of signals in the visual system, the color and brightness of the lemon, and in fact of all objects, would seem to change drastically when the intensity of the light falling on them changed. Without such processing, the lemon that looks yellow and bright on the tree might look dark and violet in the kitchen. It is only within the past 20 years that vision scientists have begun to develop a clear understanding of the nature of these processes and of their consequences for human visual perception.

For a while it looked as if the greater insights into the mechanisms of human sensory responses would come from the anatomists and physiologists. Once the technological tools (microscopes, oscilloscopes, amplifiers, computers, chemical probes, etc.) were made available to the "wet" scientists and once they had gotten thoroughly imbued with a mechanistic viewpoint, they proceeded to lay bare the immediate physical substrates to human sensory responses: the mechanical transducers (e.g., the optics of the eye, the middle and inner ear), the sense cells and the way the physical stimulus (light, sound, pressure) activates the cells, the neural pathways and their connection in the brain, etc. Spectacular successes were achieved: It was demonstrated how all of vision is funneled through a chemical stage inside the receptor's cells and that the wavelength dependency of vision could be accounted for by the absorption properties of the chemical visual pigment molecules at different wavelengths. More recently it has been shown that the individual nerve cells in the visual pathways behave in a manner more complicated than merely being photo cells. They respond best to rather specific light shapes, for example, bull's-eyes or bars of light. The excitement generated by these findings was not about the discovery that individual cells in the sensory nervous system behave in a complicated way, but that the kind of complication they exhibit matches the complication that behavioral researchers find in human and animal responses to visual stimuli.

Over the years psychophysicists had observed many kinds of visual phenomena that did not fit in with the view that the visual system operates as a simple set of light transducers. Blurred edges are seen as sharper than they actually are, a feature essential to the successful reproduction of photographs in newspapers or moving images on TV screens. The contrast of adjacent areas seems enhanced. Color values depend on brightness and color context: A piece of gray pottery appears dark on a white tablecloth but light on a dark mahogany table. Spatial patterns appear larger or smaller or even tilted in ways that are predicated on neighboring patterns, a fact that architects for centuries have known they must take into account. Many of these effects are celebrated as "visual illusions," but their more insightful investigators have always accepted them as manifestations of the transformations of sensory signals in their passage within the human central nervous system. The discovery of neural networks with properties that fit almost exactly some of the specifications set by psychophysicists constituted an important landmark.

Although there was general jubilation among sensory physiologists that they had "explained" behavior, or at least validated observations of behavior in mammals, in fact the situation was precisely the reverse: What might otherwise have appeared as an arbitrary arrangement of excitatory and inhibitory connections in the nervous system suddenly fell into place as meaningful once it was realized that it matched the performance of the whole organism.

This interaction between physiological findings in animals and behavioral observations in humans must be regarded as a high point in the investigation of sensory phenomena. Matches were sought between the ability of an observer to distinguish colors, shapes, movement, and three-dimensional depth of targets, on one hand, and the inbuilt selectivity for such features of cells in the retina and sensory brain, on the other. Cells that tend to respond predominantly to a line of a certain tilt, for example, are said to have this line as their "trigger feature," and the dissection of animal behavior into channels delineated by trigger features of cell classes in the incoming stream of information seemed to be the obvious approach. All that was needed was to enumerate fully the cells' trigger features and describe their properties in detail, and, in principle, the sensory process would be characterized.

As this process was pursued during the last decade, a difficulty emerged. It concerned the first step in the process, the enumeration of cell classes by their trigger features. Single-cell experiments in mammals are difficult and the search for the precise trigger feature of a cell is time-consuming. In fact, the possible combinations of visual stimulus parameters in the domains of space, time, brightness, and chromaticity is astronomical, making it impossible to proceed in a systematic manner through all possible target positions, shapes, velocities, colors, etc., in order to state the cell's adequate stimulus with assurance. Where anatomy helped, as in the retina or the first relay in the brain, the approach was moderately successful, but now that the search is being continued into secondary and tertiary cortical projections, results are becoming more and more ambiguous.

Thus there is a need for organizing principles to be brought into the single-cell laboratory or, if you will, preconceived ideas as to the likely combination of stimulus parameters that may constitute the trigger features of a cell. And here psychophysics reemerges as a guide. As the routing of sensory signals within the central nervous system becomes fuzzy, with many different loci of activity, the pattern may be clarified by knowing what the major response modes are of the whole organism.

Over the years steady progress has been made in the delineation by psychophysical means of the modes of signal processing by the human nervous system. A great deal of the most reliable data is obtained by threshold experiments, i.e., by the determination of the least stimulation that can be detected by the human nervous system. A threshold, as the word suggests, is a boundary or dividing point. When we speak of a light threshold we mean the energy level of the light such that above that level it is seen or sensed; below that level it is not. Threshold experiments are ones that measure this energy level for a wide variety of stimuli. By imaginative manipulation of stimulus variables, enormous areas have been charted. For example, the minimum quantity of light needed in order to be detected, called the visual threshold, is now no longer being measured just for a spot of light: The spot can be placed on backgrounds of various sizes, shapes, and colors, and in this way the confluence of excitation from neighboring regions (in time, space, and color) can be ascertained. Thus it is possible to predict whether certain light signals are visible in a whole range of road, rail, and aviation traffic situations. Instead of using spots of light, thresholds can be determined for more complex and interesting shapes--and it has been found that detection can be improved by the judicious selection of the overall shape or size of a pattern. The same applies to the velocity of targets.

A particularly interesting approach is to adapt the visual system to a particular stimulus situation by having a subject view it for perhaps a minute. When, for example, a band of stripes of given spacing is viewed for a while, other stripe spacings look wider or narrower than before, and they require a greater luminance ratio between the light and dark stripes--a higher contrast--to be detected. In finding the range of stripe widths affected by adaptation to a particular pattern, it becomes possible to describe the channeling of information in the visual sensory system. Stripes or other patterns whose visibility is unaffected are presumably not processed in the same channel; conversely, the extent to which any pattern is affected helps to outline the characteristics of the channel.

Channels--three in number--to transmit color information have been known to exist for a long time, but the description of channels used in spatial vision and in the velocity domain is relatively recent. Channels can be quite specific. For example, it has been shown that adaptation to the back-and-forth jitter of a pattern leaves unaffected the detection of zooming motion--signifying the existence of several separate motion channels. (For a detailed application of these results, see below.)

Another area in which modern psychophysics is pointing the way is in outlining the ultimate sensitivity of the organism to small changes in stimulus situations. Time discrimination in hearing can be in terms of microseconds, in vision a fraction of a millisecond. The eye can resolve scenes as well as can be done by any diffraction-limited optical instrument of the same aperture. Shades of color can be discriminated by most observers in a way that taxes the capability of colorimetric schemes. In terms of the sensitivity of the eye to light, one or two light quanta arriving simultaneously are enough to be seen; no photo cell can do better. The localization ability of the human visual apparatus is so fine--a few seconds of arc or one hundredth of a milliradian--that it is called hyperacuity to set it off from ordinary telescope resolution. All these fine discrimination capabilities of the human visual sensory apparatus betray the operation of neural circuits that transcend by a factor of 100 the current best estimates of what nerve cells by themselves can do. They therefore point to the next task of neurophysiology: the elucidation of computational machinery in the brain that accepts relatively crude sensory input and processes it, by means as yet unknown, to yield signals of exquisite refinement. How they are stored (memory) and compared with each other (cognition) is the subject of other facets of behavioral science.

Audition

Psychoacoustics, the psychophysics of auditory phenomena, owes much to two pioneers who initiated the earliest investigations in this area. Lord Rayleigh, the English scientist, almost single-handedly developed the modern field of physical acoustics and also began the scientific exploration of binaural hearing. H. von Helmholtz, the German physicist and physiologist, pioneered the first systematic analysis of auditory phenomena. Their discoveries, theories, and speculations resulted in the field of psychoacoustics, which relates what one hears to the physical properties of the sound stimulus.

An important aspect of any sound is its intensity, since as intensity is varied the loudness of a sound changes. Rayleigh in 1882 invented a disc, now named after him, which provided the first means of measuring physically the intensity of a sound field. Prior to that time the frequency of a sound was the physical parameter of major interest, because changing the frequency of a source altered the pitch of the sound, and frequency could be measured physically with some accuracy.

One of the earliest questions involved how the sensation of pitch is conveyed within the nervous system: How are the major physical aspects of sound, frequency and intensity, coded? In 1863 Helmholtz proposed what became known as the resonance or place theory of hearing. He suggested that somehow the different places along the cochlear duct of the inner ear, where the sense organs of audition reside, are differentially responsive to different frequencies. The analogy is a harp. Just as different strings of a harp vibrate "sympathetically" to different frequencies, so different places along the cochlea vibrate to different frequencies of sound. Place thereby codes the frequency of the sound, and the amplitude or vigor of responses (e.g., how many times the nerve fires per second) codes the intensity of the sound vibration at that frequency.

This theory was a natural extension to audition of the classic Young-Helmholtz theory of color vision. The eye contains three different receptor types, which are differentially responsive to long, medium, and short wavelengths. Thus the color quality is thought to be coded by which receptors are active: the vigor of their total activity codes intensity. The major difference between the place theory of hearing and the Young-Helmholtz theory of color is that, instead of three receptors, there are hundreds of different qualities (different neural fibers) representing the pitch of the sound.

The assumptions of place theory were comfortable to the physics community in part because they nicely meshed with the remarkable mathematical insights of J. Fourier. Fourier's theorem asserts that a complex periodic wave can be represented as a sum of simple sinusoidal vibrations. Place theory assumes that the ear performs such a decomposition, each place resonating to a distinct sinusoid, and the collection of places thereby representing the complex periodic sound.

Early psychoacoustics was not fully systematic because precise means of generating sound stimuli were not available. Typical sounds were tuning forks or crude sirens. Modern psychoacoustical investigation began when reliable instruments for generating and controlling the physical stimulus became available. This era began with investigators adapting the new electronic technology to the generation of sounds via headphones (invented by Alexander G. Bell in 1876) and progressed with later improvements in headphones and loudspeakers. The earliest systematic studies were carried out, not surprisingly, by the Bell Telephone Laboratories about 1930. Their interest in psychoacoustics arose because the ultimate arbiter of the quality of any acoustic transmission system is the sense of hearing. About the same time G. von Békésy, a Hungarian telephone engineer, began studying hearing for the same reason. His investigations later earned him the Nobel prize.

The earliest psychoacoustic experiments of the modern era (e.g., those of R. L. Wegel and C. E. Lane in 1928) studied the way in which a tone of one frequency could make a tone of another frequency difficult or impossible to hear. This effect, called masking, is important since, to the degree that masking is effective, communication is impossible. Information on masking has practical applications in understanding how we hear, or fail to hear, in noisy environments. Information about masking contributed in important ways to the development of effective radio communication systems in aircarft in World War II.

Gradually the model of the ear as a series of resonant channels, each responsive to a slightly different frequency, became a widespread idea and was used to explain a variety of masking results. This general notion was completely consistent with Helmholtz's earlier resonance theory. A most important related phenomenon is called the critical band of frequencies. The basic idea is that when two frequencies are sufficiently close together that they both activate the same resonant channel, then they interact in a way that is quite different from the case of two more widely separated frequencies that activate different channels. For example, consider what happens to the detectability of a pure tone in a narrow band of noise centered around the tone. At first, as one increases the width of the noise band, the detectability of the signal decreases because the total disturbance in that channel grows with the number of components present. But once the band is so wide that any increase affects different resonant channels, then the detectability does not change any further. This is one way to estimate the width of the band. Because many interesting psychoacoustic phenomena are understandable in terms of critical bands, their exact nature has been the focus of considerable experimental and theoretical work.

About the same time as the ealy psychoacoustic experiments, Békésy began to explore the anatomy of the cochlea and was eventually successful in seeing the vibration of the basilar membrane, the delicate tissue on which rest the hair cells, which are the receptor elements of the auditory sense. The amplitude of the vibration is very small; even at enormous intensities the amount of movement is barely detectable using the option methods Békésy employed. In the 1970s much more subtle techniques have been used to study these vibrations (e.g., the Mossbauer technique and laser interferometry). By and large the latter measurements completely support Békésy's earlier observations, confirming that each place along the membrane responds only to a certain narrow range of frequencies, just as place theory would have it.

Following this earlier work, a large number of psychoacoustic investigations explored the finest discrimination that could be made of a small change in a basic physical parameter of the stimulus. One could, for example, hear a change of less than 10 percent in intensity, a change of less than 3 percent in frequency, and a change in angle of spatial locus of less than 1 percent if the source was located straight ahead and emitting a broad spectrum. All of these studies of auditory acuity were aimed at inferring something about auditory processing from a measurement of the smallest detectable change. As in the visual system, many potential hypotheses about auditory processing are untenable because they would not result in sufficient sensitivity to accord with the measured sensitivity of human observers. Many and varied experiments of discrimination capacity continue to be pursued. These new experiments refine and limit the number of reasonable hypotheses concerning the details of acoustic processing.

Animals other than humans have also been studied because in some cases their sensitivities exceed those of humans. As their response to special ultrasonic whistles demonstrates, dogs and cats can hear higher frequencies than can people. More remarkable still, bats navigate and catch prey using the reflections from the ultra-high-frequency pulses that they produce. Bats used the principles underlying sonar and radar millions of years before humans discovered them.

Meanwhile, physiologists continued their exploration of the hearing mechanism in an effort to understand the details of auditory processing. A major breakthrough was the ability to record, electrically, from a single auditory fiber in the acoustic nerve, the VIII cranial nerve bundle. These recordings revealed that each fiber is maximally sensitive to only a narrow band of frequencies, the width of the band increasing with the intensity of the tone. In effect the fiber acted very much as a resonance filter, maximally sensitive at one frequency but capable of responding to other frequencies if the intensity is sufficiently large. In a plot of intensity versus frequency, the curve separating the region of responsiveness from that of no response is called a "tuning curve," and the measurement of a fiber's "tuning curve" is now an essential first step in any serious study of the auditory nervous system.

The tuning curve tells us which place along the basilar membrane contains the hair cell that drives this fiber. Tuning curve analysis is, in short, completely consistent with Helmholtz's place theory. In recent years psychophysicists have devised a masking experiment that generates a resonance-like curve closely resembling the physiologically measured tuning curve. In addition, near the edges of the tuning curve one finds frequency and intensity combinations that appear to produce suppression, a process that can cancel the effects of masking. Again, both physiological and psychophysical data show strong similarities. The observation of suppression suggests a region of inhibitory action flanking an excitatory center, reminiscent of a lateral inhibition mechanism such as that found in vision and the skin senses (see the section on visual modulation transfer function and visual disorders).

Although the place-resonance theory of Helmholtz is supported by all of the available peripheral data, it is not the whole truth. The perception of pitch, especially the pitch produced by complex periodic stimuli such as those arising from musical instruments, is not as simple as the original theory would have it. About 1940 J. F. Schouten of the Netherlands reported some experiments in which subjects heard a low-frequency pitch, called the residue pitch, as a result of certain combinations of high-frequency components with no energy whatsoever at or near the perceived pitch. Thus, although activity at one place signals the corresponding pitch, that same pitch can be produced by the combination of activities at other places. Moreover, as several experiments demonstrated, the low pitch is not the result of nonlinear distortion. This is an auditory illusion in the same sense that there are visual illusions. One is hearing something quite different from what is in fact present in the stimulus. And as with visual illusions, it is important for two different reasons. First, it places a very strong constraint on theoretical ideas about how the auditory system works. Second, it is an illusion that can affect the practice of engineering acoustics. For example, noise engineers were initially baffled by the fact that people complained of noise at low frequencies in jet engines when there was little or no energy at these frequencies. The fan blades were creating very regular patterns of energies at higher frequencies that created significant residue pitches. A decade of further work, both in this country and in the Netherlands, has further clarified and confirmed Schouten's original observations concerning residue pitch. The current consensus is that residue pitch undoubtedly represents the action of some more central process interpreting and integrating the peripheral sensory information. The recognition and acceptance of the facts of residue pitch have also been important in advancing understanding of nonlinear phenomena in hearing. Such nonlinear effects are important and ubiquitous but were largely misunderstood by early investigators.

In summary, exploration of the auditory sense has been one of mutual support among investigators working in physics, physiology, and psychoacoustics.

Cognitive Factors in Psychophysics

Despite the highly successful interplay of psychophysical and neurophysiological research, there are major psychophysical phenomena, some of which were only fully recognized in the past 30 years, that have completely eluded physiological clarification. These have to do with phenomena of the central nervous system--sometimes called information or cognitive processing--that are only partially influenced by the peripheral information arising from the physical signals. Among the topics in this area are: the ability of a subject to attend differentially to aspects of the stimuli impinging on him or her; the trade-off that exists between failing to detect a signal and falsely responding that it is present when in fact it is not; the trade-off that exists between accuracy of performance and the time it takes to respond; how varying a signal or its surrounding environment affects the subjective growth of sensation (e.g., what does it mean to say that the noise level has been reduced by half?); and the inability of most people to identify correctly more than about seven signals that differ along just one physical dimension.

We elaborate several of these examples. Consider driving on a lonely road. One is continually scanning for danger signals--another car, fixed obstacles such as trees or rocks, and of course pedestrians. Having selected a speed at which to drive, there remains another variable under one's control--how reactive to be to apparant signs of danger. If one is very reactive, applying the brakes at the first partial indication of danger, then frequent false alarms result (i.e., braking when there is no obstacle) but less danger of striking something. If one is less reactive (i.e., waits until a clear danger is present before applying the brakes), then fewer false alarms may result but the chance of an accident rises. Clearly, the driver has available some freedom to decide just how reactive to be, both in terms of the amount of evidence collected in a fixed time that is sufficient to cause braking and in terms of the amount of time to delay before making the response. Such trade-offs are ubiquitous in sensory psychology. Considerable work has gone into their study, and rather elaborate mathematical models have been developed in an attempt to capture the basic principles of the decision processes that are involved.

This work has shown that questions such as: "How fast can a person respond to a signal of such-and-such a character?" or "How likely is it that such a signal will be detected in a certain environment?" are either meaningless as formulated or, at the very least, require very subtle answers. How fast one responds depends greatly on how many false and/or anticipatory responses are permitted; how likely one is to detect a signal depends on how many false alarms are tolerable. With exactly the same signal conditions and the same subject, the likelihood of detecting a signal can be varied from nil to certainty simply by varying how likely a signal is to occur in a fixed time period or the nature of the rewards for correct responses and the punishments for the two types of errors. The past several decades have provided us with much data and sophisticated models concerning the nature of this tradeoff.

Loudness grows with the amplitude of the sound wave, visual size with physical extent, and the sensation of shock with voltage. But exactly how do these sensations grow? This question was initially raised during the 19th century; G. Fechner's attempt to answer it led to the beginnings of psychophysics, and it has been inextricably intertwined with all subsequent theoretical developments in the field. One of the more striking of these emerged in the 1960s from results obtained by S. S. Stevens, who simply had subjects assign numbers to signals in proportion to the subjective sensations they engendered. It turns out that they can do this seemingly impossible task with considerable regularity. As a first approximation, sensation measured this way grows as a power of signal amplitude, A^β, where A denotes amplitude and β is an empirical exponent. The exponent involved varies from less that the cube root (loudness and brightness), through the linear function (line length), to something in the neighborhood of the cube (electrical shock). There is much work going on attempting to understand the theoretical basis of these relations, to understand how various factors affect these subjective sensations, and to understand how they combine.

Some specific questions are these. Suppose that a sound is composed of several pure tones. Can the loudness of the combination be predicted from the separate loudness of the components? From what was said in the section above on audition, it comes as no surprise that the answer is much affected by the structure of the critical bands.

Given separate sounds to the two ears, how does the overall sensation of loudness depend on the individual loudness in each ear? Consider the apparent size of the moon. We are all familiar with the fact that the moon rising over the horizon seems huge compared with its size when it is high in the sky; subjective estimates yield a factor of about two. Yet from photographs or by viewing the moon through a mailing tube we know that the visual angle at the eye is virtually identical in the two cases. So, why does it seem larger at the horizon? No one really knows, despite the fact that the phenomenon has long been recognized and a number of attempts have been made to explain it. Such an illusion, one of many of which we are aware, is surely not an isolated curiosity; rather it tells us something about the basic information processing carried out by the brain on the input signal.

Another phenomenon that has been recognized since the 1950s but whose basis is still not fully understood is this: If one is asked to identify which of two sounds, one twice as intense as the other, has been presented, one can do so with perfect accuracy; sounds separated by that amount are never confused. Suppose the total number of sounds is increased to 10 and successive ones are still spaced by the same factor of 2; errors will then be made, e.g., sound 8 will sometimes be called sound 7 or 9, despite the fact that all can be perfectly discriminated as pairs. The data tell us that the ability to identify sound 8 and to discriminate it from 9 depends on other sounds that might be presented. The same is true of brightness and most other modalities. (The major exception is auditory frequency: Some people exhibit the phenomenon of almost perfect pitch.) As far as anyone knows, the brain has exactly the same peripheral information when a particular signal is presented, whether it is 1 of 2 or 1 of 10, yet once the number exceeds about 7 there is increasing difficulty and confusion in identifying which signal is presented. Why? Various theories have been offered, but to date none seems fully adequate. It is clear that as yet we do not fully understand the nature of the coding involved and the processing done by the brain.

We mention these easily demonstrable conundrums because they are both very familiar and very difficult to understand. There can be little doubt that the brain, when processing even the simplest of signals, is using information that extends well beyond the particular signal at hand. Psychophysicists are making attempts--often mathematical ones--to formulate the types of processes that may account for these phenomena.

Skilled Performance

Another important area of basic research is the development of quantitative models for how a person uses visual or auditory information in combination with his or her cognitive skills to control and manipulate machines.

When one drives a car or flies an airplane, the senses, particularly the eyes, take in information from the environment, and the brain processes the information, together with goals or intentions, to send signals to activate the muscles. The muscles in turn move the steering wheel or control stick in order to control vehicle movement. When a person serves as a vital link in a control system, it becomes very important to be able to understand the human behavior involved in engineering terms. The sensory-motor response of the driver or pilot affects the system's overall stability and performance, just as do the tires or ailerons.

To the naive observer the task of driving seems to be automatic and to take very little mental effort. In fact, much of the activity involved appears to be unreportable. The individual performing the control task cannot describe how he or she does it. Research has shown, however, that it does require mental effort and, in fact, in most situations such control employs processes that would be described as distinctly cognitive.

A pilot following the glide slope needle that directs the plane toward the runway is performing the simplest of tracking tasks. However, the choice of a control action entails selecting a sequence of muscle commands that must take account not only of the stiffness and inertia of the pilot's own arm but also the sluggish dynamics of the plane itself. The response of the plane is so subtle that one cannot depend on visual cues or pressure on the seat of one's pants to provide adequate feedback. One must integrate these cues with a learned prediction of how the vehicle will respond. We say the pilot has an internal model of the vehicle dynamics and incorporates its response into his muscle command planning.

This point is made forcefully when a tire blows out or a yaw damper fails and the handling characteristics of the vehicle change suddenly. The control behavior appropriate to the system prior to the emergency is quite different from that required after the change. The internal model must be changed and changed rapidly.

When a pilot executes a preplanned maneuver such as a turn to final approach, the cognitive demands are even greater. Since no pathway in the sky exists, the basis for executing a particular turn must be drawn from memory. Some argue that a particular maneuver comes from a schema representing turns in general made particular to the set of conditions found when the turn is begun. The execution also must take account of the specific vehicle characteristics and the nature of the controls involved.

The first attempt to represent human control behavior in such engineering terms was accomplished in 1947 by A. Tustin, a British scientist. The first extensive set of data that described tracking behavior in terms of control engineering equations was completed in 1956 by J. I. Elkind. That work represented a significant advance in experimental measurement of dynamical systems as well as a landmark contribution to the literature in engineering psychology.

Since that time modern control theory has extended the power and complexity of the systems that can be analyzed. The optimal model of manual control describes the behavior in terms of state estimation, information processing, and response generation. It is interesting that the model also assumes that the driver or pilot employs an idealized internal representation of the system being controlled, exactly analogous to the cognitive models described in an earlier section.

Research on manual control and the prediction of human motor performance still has difficulty predicting the effects of learning on performance. There are special difficulties in predicting those practice effects associated with voluntary maneuvers for which there is no explicit pattern to be followed. Manual control modeling research is also being broadened to include prediction of the behavior of multiperson crews when the task of actually controlling the vehicle is overlaid with a myriad of decision-making and procedural tasks, sometimes referred to as supervisory control. One such model describes the activity of the three-person crew of a commercial jet transport during approach and landing. Others have been designed for bicycle riding and motorcycle handling.

Besides the conceptual understanding they contribute, manual control models have been practically useful, for example, in predicting critical design conditions that need careful experimental study. The models of jet transport crews were developed to evaluate alternative landing procedures and staffing requirements.

As this discussion reveals, research on manual control has provided a most interesting and stimulating interchange between experimental psychologists and engineers. It has captured the creative talents of researchers from several fields, not only in terms of the development of new theoretical concepts and measurement procedures, but also in translating these concepts into data and methods useful in the system development process. When analytic methods are available to predict performance, it is possible to narrow the range of candidate designs that need to be evaluated in depth. While it would be impossible to estimate the cost savings attributable to these developments, it is clear that they have substantially reduced the development time and the costs associated with experimental evaluation of alternative designs.

EXAMPLES OF THE INTERPLAY BETWEEN BASIC RESEARCH AND APPLICATIONS

In discussions of the relative importance of basic and applied research programs, an economic analogy is often made. Basic research is the savings account, to be accumulated and husbanded against the lean years, when the checking account of applied research runs low. The cash reserve, it is said, is drawn upon to replenish the drained resources of the experts in application.

The analogy is incomplete in this sense: It neglects the vital presence of those whose activities move the masses of information in and out of the fund of knowledge. Basic knowledge can no more be static than the funds in a savings bank can if there is to be any gain from it. But a requisite for moving, altering, and adding to information is people who are proficient in handling and remodeling it. What William James spoke of as the cash value of an idea is tangible only in the hands of a skilled barterer.

We turn now to this interplay of basic and applied research. Several examples of applications, one inexorably intertwined with the basic research, have already been mentioned. In what follows we take up a number of additional cases in which basic psychophysical research led or is leading to significant applied work. In some cases, the application is complete. In others, the basic work has suggested an idea for solving a problem and the work is currently being pursued. As always, promising routes seem so until they fail; few roads end at the goal, and most are a lot rougher and not nearly so straight as anticipated.

Two Modes of Visual Processing and Night Automobile Accidents

For some time psychologists studying visual perception have posited two independent and dissociable modes of processing visual information. The focal mode is concerned with object discrimination and identification or, more generally, the question of what. It is subserved primarily by the cortex and is typically well represented in consciousness. Because focal functions involve the higher spatial frequencies, i.e., finer visual textures, they are optimal in the central visual field and are systematically related to both luminance and refractive error. The other mode of processing, referred to as ambient vision, is concerned with spatial orientation or, more generally, the question of where. The properties of the ambient and focal modes differ along many dimensions. Although spatial orientation is certainly possible, if not superior, with the central visual field, it is adequate with stimulation of the peripheral retina in spite of the coarse resolution properties of the latter. Coarse patterns are sufficient for ambient functions, so they are less sensitive to both refractive error and luminance than are focal functions. With respect to consciousness, the ambient system is often poorly represented, although by directing attention one can be aware of ambient activity.

A number of recent ablation studies, as well as observations of brain-damaged people, have suggested that it is possible for some orientation ability to be spared despite loss of focal vision. L. Weiskrantz has referred to this interesting phenomenon as "blindsight." For our purpose the fact that it is possible to walk while reading demonstrates the dissociability and some basic characteristics of focal and ambient functions. Even though attention is dominated by the reading material, orientation in space is carried out confidently and accurately by the peripheral visual fields operating at an unconscious or subconscious level. If illumination is lowered or the retinal image blurred, the ability to read is degraded but orientation is relatively umimpaired.

A critical problem in vehicle guidance, which can be understood in terms of the theoretical approach, is the high frequency of nighttime driving accidents. Automobile accidents, of course, have multiple causes. The role of illumination is demonstrated, however by studies indicating that, when other factors are held constant, accidents, particulary those involving collisions with pedestrians, increase dramatically under lowered illumination. It is well known that under twilight and nighttime conditions many visual capacities, such as spatial resolution, stereoscopic depth perception, contrast discrimination, and reaction time, are degraded. This is reflected in analyses of nighttime accidents in which drivers frequently report that they did not see a pedestrian or other obstacle in time to stop. In some cases, the sound of impact was heard before the driver was aware of the pedestrian. What is curious is that drivers typically do not reduce their speeds at night, even though they are probably aware through personal experience, or even through knowledge of literature, that their vision has been degraded.

A possible explanation for this paradox may be derived from considering the two modes of processing. Driving an automobile, like walking, flying, and sailing, involves two parallel tasks. Spatial orientation is accomplished by steering the vehicle, which requires continuous evaluation of the location of the vehicle relative to the road. In terms of the two modes of processing, steering is concerned with where and is an ambient function. Driving also involves focal vision, the role of which is to monitor the roadway ahead for pedestrians, other vehicles, and obstacles, to read traffic signs and monitor signal lights, and to judge the distance and speed of other vehicles. In daylight both the ambient and focal modes are operating at their maximal capacities. Under twilight and nighttime conditions, however, there is a selective degradation of the two modes. Focal visual functions are degraded, i.e., spatial and stereoscopic acuity are reduced and contrast sensitivity is diminished. (For many individuals the ability to appreciate detail is further degraded by a condition known as night myopia.) The efficiency of ambient visual functions, however, is not reduced by lowered illumination. As long as minimal visual stimulation is available, it is possible to steer the vehicle adequately. In terms of the performance information available to the driver, it is the ambient mode that dominates. Since the demands on focal vision are intermittent, information about the degradation of focal vision is only rarely reflected in the operator's performance. As a result, the driver is not aware that there is a problem with the degradation of focal vision and therefore typically maintains the same velocities at night as during the day.

As is often the case, understanding the basic cause of a problem suggests methods for amelioration. To reduce the high nighttime accident rate a number of possibilities are apparent. Obviously, illuminating highways would be expected to be effective, and this is supported by empirical observations. However, economic considerations limit this possibility. Other alternatives are to post different maximum velocities for nighttime and daytime driving conditions. Before the introduction of the uniform national 55 mph speed limit in the United States, only a few states followed this practice, usually on major highways. To our knowledge, different speed limits have not been posted in areas where degradation of focal vision would be expected to play a role in accidents involving pedestrians. Another possible measure is to educate drivers regarding the potential dangers associated with the selective degradation of vision at night. This procedure would be expected to be particularly effective for younger drivers, whose habits have not been established. The implications of selective degradation should also be communicated to pedestrians and cyclists, who should be encouraged to take special measures to increase their visibility at night in order to compensate for the loss of focal vision of drivers.

Visual Modulation Transfer Functions and Visual Disorders*

A powerful tool for studying visual processing is the measurement of what is called visual describing function or, more loosely, visual modulation transfer function (MTF). Briefly, the procedure for studying the visual MTF is like this: A person looks at a television screen on which a pattern of stripes is present. The experimenter determines how much contrast the stripes must have in order for the person to be able to detect them. This just-visible contrast depends strongly on the spacing of the stripes, on whether they are flickering and if so at what frequency, and also on the colors of the stripes and the condition of the subject's eye.

---

*The authors express their thanks to Jane E. Raymond for valuable suggestions for this section.

In general, measures of the MTF provide significant amounts of information about the general behavior of the human visual system and, together with data from physiological experiments on the eyes and brains of various animals, are rapidly leading us to a good understanding of the anatomical and physiological structure of the human visual system. To give just one example, it is now clear that certain neural structures in the human retina are organized in such a way that when a small spot of light falls on the retinal surface, it not only produces an increase in the activity of a few nerve fibers leading from the eye to the brain, but it also produces a decrease in the activity or responsiveness of all of the nerve fibers corresponding to adjacent spots on the retina. This is called lateral inhibition. We have good estimates of how strong these effects are, the distances across the retina over which they operate, their sensitivity to the timing of changing light intensities, how these factors vary at different parts of the retina, and many other similar parameters.

This kind of information is important in a number of ways. First, and perhaps most important, it provides strong hypotheses about how the brain itself works, because the retina is closely related to much of the brain in its embryology and structure. Second, it provides explanations for many visual phenomena, such as the fact that objects look the same regardless of the intensity of light illuminating them, which has puzzled scientists for hundreds of years. Third, it permits accurate predictions of the appearance of unfamiliar patterns of light, such as those experienced by astronauts when traveling in outer space.

The use of the MTF in visual psychophysical research has also been valuable in advancing our understanding of the perception of complex patterns such as letters and faces. Any complex pattern can be broken down into a series of component stripe patterns, each with a specific width, orientation, contrast, and position. By viewing the visual system as a series of filters, each sensitive to a different set of stripe widths and orientations, accurate predictions concerning the perception of complex objects can be made once the component stripe pattern used to construct the complex pattern is identified and the MTF of the visual system is measured. This approach has been valuable in basic research on object perception and recognition, and it also has been successfully applied to practical issues.

Although the concepts surrounding the visual system MTF have been useful in areas such as aviation engineering and electronic visual communications engineering, the most widely explored applications of the MTF to date have been in the area of clinical medicine. Three examples follow.

Contrast sensitivity functions were first used clinically by neurologists investigating patients with disorders of the central nervous system such as cerebral lesions, epilepsy, and multiple sclerosis. As a diagnostic tool, measurement of the MTF has been particularly successful in detecting visual involvement of multiple sclerosis in patients who appear normal on ophthalmological examination. Aside from the diagnostic value of this procedure, the data obtained can also be employed to indicate a physiological rather than psychogenic basis for visual complaints and perceptual difficulties.

Second, it has been known for centuries that those children with eyes that do not point in the same direction exhibit characteristics different from those who need glasses but have not worn them. Differences are also found in their MTFs, suggesting differences in the mechanisms of visual loss. Study of the differences may lead to improved methods of prevention.

A third example of the application of MTFs to visual pathology is in the detection of glaucoma. During 1979 and 1980, it was shown that the MTFs for glaucoma patients and many glaucoma suspects differ in characteristic ways from those of people with normal vision. These differences may provide an important means of detecting glaucoma before it has caused serious visual loss. Equally important, the nature of the MTF differences between glaucoma patients and people with normal vision provides important evidence about the actual pathological processes that occur during the course of the disease.

In summary, the study of how the visibility of a pattern of stripes is affected by various characteristics of the stripes and by properties of the eye is leading us to a better understanding of the nature of both normal and pathological human vision.

Speech Perception and Reading Machines

Our understanding of speech perception, as well as our ability to put that understanding to practical use, derives from the confluence of several currents in science and engineering. We trace one such current here, one that exhibits the interplay of a practical problem and the basic research in psychophysics that contributed to its solution. The example is striking, because the practical problem seemed at the outset to require very little more knowledge than was available at the time and also because the basic research, once undertaken, has proved useful beyond the demands of the particular problem that provided the initial stimulus.

The problem, first seriously attacked at the end of World War II, was to build a reading machine for the blind, a device that could scan print and produce an understandable acoustic signal. Some, perhaps all, of those who set out on this enterprise were guided by the assumption that the device had only to produce, for each letter, a pattern of sound that was distinctively different from the pattern of all other letters. Blind users would presumably be able to read after learning to associate the sounds with the letters. The rationale for this device was to be found in an obvious fact and in a seemingly obvious assumption about that fact. The fact is that sound and the ear work well together for the purpose of conveying in speech the very phonemes that the letters of the writing system (approximately) represent. The seemingly obvious assumption was that the sounds of speech are related to those phonemes in a straightforward way, much as the sounds of the reading machine would be related to the letters, and that, in this crucial respect, perception of speech was assumed not different in principle from the perception of any other sounds. Accordingly, it was expected that the sounds of the reading machine would work as well as speech, provided only that they were distinctive and that the users were given sufficient training.

In fact, no arbitrary sound alphabet could be made to work well, no matter how distinctive its individual elements or how long the training of the users. As the conclusion took shape, some of those engaged in the undertaking began to suspect that the assumption about speech and its perception was wrong. There is apparently something special about speech, something that the arbitrary sounds of the reading machine could not capture. Thus it happened that some investigators put aside their work on the reading machine in order to take up the basic research on speech perception that was required if they were ever to find out in what ways speech is special.

At the outset, the task of studying the perception of speech was no different in principle from that of studying the perception of anything. The first step is to find the cues--the physical stimuli--that control the perception; more generally, of course, the aim is to characterize the nature of the relationship between those cues and the precepts they support. Nor is the research procedure different in principle from that which had always been followed in perceptual psychophysics. It was to use methods of analysis to formulate working hypotheses about the cues and then to test those hypotheses by synthesis--that is, by manipulating the physical signal in ways appropriate to the hypotheses so as to determine the effects on the sound as perceived.

But applying the standard procedure to speech was complicated by several special difficulties. For one, many of the most important acoustic cues for speech are in the dynamic aspects of the acoustic pattern, a circumstance that imposes special requirements on the development of an appropriate research synthesizer. A further difficulty lay in the fact, not fully understood when the research began, that the relationship between acoustic signal and phonetic percept is peculiar in ways that make it significantly harder to see just what form the cues may take and where in the signal they may be found. As a result a considerable amount of time had to be spent in designing and perfecting an appropriate research synthesizer. The synthesizer needed to provide control of the putatively relevant aspects of the signal, including especially those of a dynamic character, and also needed to be sufficiently convenient so as to permit the very large amount of experimentation that proved necessary to disclose the special nature of the relationship between signal and phonetic percept.

As the research progressed, many acoustic cues were found and their effects evaluated. With that done, it was possible to see some of the general characteristics of the relationship between the cues and the percepts. Perhaps the most important of these is that the acoustic segmentation does not correspond, as had been originally supposed, in a straightforward way to the segmentation of the phonetic message. Rather, the information about any particular phonetic segment is widely distributed through the signal and overlapped, often completely, with acoustic information appropriate to other segments, thus reducing the number of acoustic segments per unit time that must be perceived. A simple but effective way to demonstrate this high degree of interdependence is to try to substitute identical phonetic elements from other contexts. For example, suppose one tries to synthesize on a tape recorder the word cat by splicing the c from car, the a from has, and the t from cut. The result is unintelligible gibberish. This is so because, in production, the phonetic segments are coarticulated, with the result that information about several successive segments is normally encoded into and transmitted simultaneously on the same parameter of the acoustic signal. To disentangle those segments in perception requires a specialized process. But if one possesses that specialization, as humans do, then the parallel transmission of segmental information that characterizes the speech code makes it a uniquely effective way of communicating language by sound.

Once the nature of the speech code was understood, it became possible to synthesize speech by explicit rule and thus to have, in principle, an important component of a reading machine that produces, not arbitrary sounds, but speech. To synthesize speech by rule means that, beginning with an input string of letters (or, more properly, their phonetic transcriptions), one produces speech by automatically applying explicit encoding instructions of a kind that can be dealt with by a computer. As first produced in 1958 (though not then by computer), speech synthesized by rule was at best intelligible only in the narrow sense of the word. Listening to it took an effort. Indeed, the extra attention needed to perceive the phonetic aspect of the message was often such as to defeat the attempt to grasp its meaning. But progress continued, and now the speech that can be synthesized by rule is much improved.

A reading machine for the blind is rapidly becoming a reality. Such a machine requires, in addition to rules for synthesis and the means for implementing them, an optical character reader to identify the printed letters and a method for converting English spelling to the phonetic transcription that keys the synthesis. These components exist and can be appropriately linked, so a blind user has a way to convert print to reasonably intelligible speech. One may wonder, of course, whether the devices currently available are good enough for all purposes. Do they, for example, still require too much effort of one who wishes to read large quantities of difficult text? Further research and development will almost certainly produce further improvements.

Automatic synthesis of speech has application not only to the blind but also to some people who are, for any of several reasons, unable to talk. In business, industry, and education, speech synthesis is finding increasing application in human-machine interactions of various kinds. Moreover, the knowledge gained through research on speech perception has proved useful for purposes other than the synthesis of speech. It provides essential information for attempts to build automatic speech recognizers, and, in particular, for understanding the nature of the difficulties that must be overcome if such machines are ever to be as versatile as we would like them to be. It furnishes the key to understanding why it is that beginning readers find it so difficult to develop an explicit awareness of the way that words can be segmented into the abstract units that the alphabetic letters represent. It helps us understand the particular nature of the difficulties that the hard-of-hearing and the brain-damaged have in perceiving speech. It creates possibilities that would not otherwise have existed for research into aspects of the biology of speech and the presence (or absence) of the appropriate biological predispositions in infants and children. And, most generally, it enlightens us about the organic connection of speech to language.

Tactile Communication for the Blind

Although braille has been used since the 19th century as the primary reading method for the visually handicapped, it requires several years of careful schooling in its use and demands that a publishing organization exist to convert printed matter to braille type. Moreover, fewer than 20 percent of the sightless population actually acquire a truly literate skill in braille, i.e., read the equivalent of four or more books a year. The skilled blind person is therefore dependent on the publications efforts of an industry having limited resources for all reading matter.

Far more satisfactory would be an aid for the blind that permits the sightless person to "read" ordinary ink print by substituting the sense of touch for the missing visual sense. Were such a direct ink-print reading system available, the sightless person would be much more independent, perhaps able to read even the labels on prescription bottles or phonograph records. One such device is called the Optacon, a contraction of optical to tactile converter.

The story of the development of the Optacon illustrates the needed interplay of knowledge, intellectual curiosity, and inventiveness. Success with the use of the tactile sense as a substitute sensory channel for the handicapped had been only moderate by the middle 1950s. At about that time, Geldard and a number of his students had well under way a program of basic research in skin sensitivity that reached a watershed with the construction of a formal skin language called vibratese. Devised as a simple alphabetic code, people could process it at rates of up to 38 words per minute after a relatively brief training period. This research effort and its publication accelerated the efforts throughout the world to use the skin as a substitute for sight.

At the same time, a program was under way at the Massachusetts Institute of Technology for developing and improving sensory aids for the handicapped. One of its engineers, J. Bliss, moved to Stanford University and there engaged in collaborative research with J. Linvill, whose daughter was severely visually handicapped. The availability of a new technology for electronic circuit fabrication emerged in the early 1960s along with an extremely small and efficient device for transducing electrical signals to mechanical motion. With these and other technical advances, which permitted the design of a very compact processing device, along with the basic psychophysical information found in the current literature on cutaneous sensitivity, Bliss and his colleagues conducted a series of research studies with support from various government agencies. Based on the skin's ability to decode pattern information in real time, they arrived at specifications for a feasible processing device.

The program of research, which took about a decade for its fruition, culminated in the production of a prototype of the Optacon, which was tested by Linvill's daughter under various conditions and with a variety of print styles and sizes. She became one of the first blind persons in the world to read ordinary printed matter at speeds of up to 80 words per minute.

Hearing Aids

Despite the personal suffering and the loss of productivity arising from hearing impairments, and despite the great advances in technology that permit the realization of almost any signal processing scheme in a cosmetically acceptable hearing aid, present-day hearing aids are still quite inadequate for a large fraction of those with impairment. Roughly speaking, current aids provide nothing more than amplitude amplification that varies with frequency. This is adequate for the impairment, called conductive loss, that is characterized by poor conduction of the acoustic energy into the neural code. Such frequency-dependent amplification is not, however, at all adequate for impairments that involve malfunction of the sensorineural processes of the inner ear or auditory nervous system, the so-called sensorineural losses. These losses result in a degraded ability to resolve different acoustic stimuli, an increased susceptibility to background noise and reverberation, abnormal changes in the loudness of a sound with change of intensity or of duration, various types of sound distortions, and internally generated sounds. When these are the symptoms, no simple frequency-dependent amplification restores normal or even functional hearing. In case of severe loss, amplification may make it possible for the impaired person to know someone is speaking but still not be able to understand the speech.

Our existing knowledge about impaired auditory perception is not adequate to permit us to characterize the nature of the aids needed to correct for sensorineural impairments. Current work on this topic should in due course lead either to the development of improved aids or to a very clear understanding why such aids cannot be developed with existing technology.

Many new signal-processing schemes are now being explored as aids for the sensorineural impairments. Amplitude compression--the systematic change of naturally occuring amplitudes to a different range of amplitudes--appears to be promising for individuals with a reduced range. Frequency compression is also being explored as an aid to people who can hear only a limited range of frequencies. In both schemes the central idea is simply to recode the original information so that it is presented in those regions of frequency and amplitude in which residual hearing exists. Other schemes explore the use of multiple microphones to simulate our binaural hearing and thus reduce the background interference. Such aids may be particularly useful for individuals with a reduced capacity to comprehend signals in complex acoustic environments. Even more elaborate aids based on automated speech-recognition or speech synthesis systems are currently being investigated.

We know that the intelligibility of speech can be improved for many impaired listeners by transforming the speech signal in ways other than frequency-dependent amplification. For example, speakers can learn to talk in special ways so that a particular impaired person is able to understand more of what is said. They neither speak more loudly nor more slowly, yet they are better understood. Studies are presently under way to try to understand how these improvements are achieved.

In general and independent of the type of impaired hearing, signal processing must be devised that enables such people to make considerably increased use of their residual hearing. The minimum criterion of success is the ability to understand speech. To design appropriate processing, it is clearly necessary to characterize in considerable detail the nature of the several types of impairment that occur, and this may very well require a substantial amount of psychophysical and physiological research. How much research is difficult to say, since we only partially understand what it is we are searching for. Success will mean both the systematic development of improved auditory prosthetics and a contribution to improved audiological diagnosis and to increased understanding of normal auditory functioning. The latter plays an important role in the improved design of high-fidelity music equipment, more efficient telephones, and better design of acoustic systems that must be used in adverse conditions, for example, military applications.

Tactile Communication for the Deaf

We end this list of applications with a final example--the use of the skin as a substitute for hearing--where success has not yet been great and we are still awaiting a breakthrough. But hope is high, and a number of investigators are pursuing a variety of different paths. We outline some of their hopes here.

Some hearing loss is so great that there is no residual hearing to draw upon. In this case, the only possibilities for restoring the function of hearing involve sensory substitution: Sound must be transformed from acoustic vibrations to either visual or tactile patterns that are then perceived.

Although it is possible to learn to understand speech with reasonable accuracy by observing the lips of the speaker (lipreading) or by placing a hand on the speaker's face (the Tadoma method employed by some deaf-blind people), both of these methods have serious limitations: Lipreading requires adequate lighting, Tadoma requires direct physical contact, and both require proximity of speaker and "listener." For these reasons, attempts at more satisfactory technology have and are being made.

The first attempt to develop a tactile aid for speech reception, more than 50 years ago, used a single vibrator, acting as a loudspeaker. Hardly more than a placebo effect was noted, i.e., the deaf persons responded to the interest of the experimenter by working harder and giving improved performance through greater effort, but better information processing was not achieved. The basic bottleneck is that the frequency range that the ear appreciates is some 40 times as wide as the tactile range of 5 to 500 Hz. Obviously, some rearrangement or recoding of the signal must be provided to give the skin the perceptual span processed by the ear.

One class of devices arranged for a set of electrical filters that would "fan out" the speech frequencies over the skin so that a particular site, when stimulated, always represented a particular frequency band. Again, this met with limited success.

Another possibility was to follow the lead pioneer by Békésy. In the course of his work on the mechanics of the inner ear, he analyzed the mechanical and hydrodynamic properties of the cochlea. He calculated the values of the mechanical constants of the membranes that supported the sensory receptors of the ear. From these computations he built a dimensional model of the cochlea that, he claimed, vibrated in the presence of sound energy in the same manner as do the receptive tissues of the inner ear. He was able to demonstrate this correspondence by having observers place their arms on the vibrating "membrane" of the mechanical model and feel the change in frequency, loudness, and location of "sound" as he manipulated these variables by electrical or acoustic means.

A prominent German investigator, W. D. Keidel, decided to use the Békésy model, which had been designed purely for the purpose of advancing the understanding of the hearing process, as a practical, tactile speech-analyzing device. Because the skin accepts only low-frequency sounds, the speech frequencies must first be reduced to the proper range for the model to work. This was done by tape-recording the speech and replaying it for the model at a very low rate, which not only reduces the frequency to the proper values but also stretches time so that one word may take several seconds to feel. Nevertheless, a few subjects did learn to understand some speech thus translated, and the investigator was encouraged to search for ways of getting speech sounds processed by computer and presented to the model so that "listeners" could feel speech patterns as they were being generated. A computer program was automated to scale the speech frequencies down to range suitable for the skin and at the same time clipped and joined the speech segments smoothly to produce a low-frequency version of the frequency-time relations of speech that lasted no longer than the original. Only limited tests have been made thus far, but this approach still seems to encourage further exploration.

Some positive results have been reported by other research groups studying tactile speech displays, which incorporate a variety of approximations to the spectral analysis performed by the ear.

We still need to determine the ultimate limits of tactile sense for speech communication, to clarify the principles governing the effectiveness of tactile displays, and to develop practical aids capable of functioning at a distance in real-world environments. In addition to contributing to the development of prosthetics for the deaf and deaf-blind, research in this area will provide increased understanding of the tactile sense, of speech perception, of design principles for sensory display, and of sensory substitution and human plasticity.

SUMMARY

This paper has reviewed what has been learned about the human senses, together with the relevant cognitive and motor components, by applying the scientific principles of the discipline called psychophysics. This information has been acquired, mainly in the past 130 years, by adapting and applying the methods of physics to problems of human beings interacting with various physical environments, ranging from very simple energy changes to complex dynamic systems, such as flying an aircraft.

As a result of these studies we now have a detailed and often quantitative understanding of many perceptual phenomena. The human senses and their associated cognitive and motor components are not passive systems reflexively reacting to the incoming stimuli, but are active transformers and processors of the applied stimulus. A great deal is known about how many of these transformations operate and their effects on perceptions.

We have also dealt with how this knowledge has been used in an attempt to solve, or at least understand, a number of practical problems. Our emphasis on sensory impairment arose not only because of the humanitarian component, but also because the best way to demonstrate a detailed understanding of some system is to be able to repair or otherwise ameliorate a defect in that system.

Success measured in these terms has been only partial, but in many areas great promise is evident. Our examples have exposed the false dichotomy between applied and basic research. Rather, as the section heading of these examples implies, there is an interplay between basic knowledge and the information and understanding gained by the attempt to apply that knowledge to concrete problems.

Reading as a Cognitive Process

Patricia A. Carpenter and Marcel Adam Just

INTRODUCTION

Reading is a central skill in a technical, democratic society, such as our own, which requires reading skill for employment, education, communication, and everyday functioning. An illiterate person cannot use a newspaper as a source of information or entertainment, cannot fill out a job application unaided, cannot understand the directions on a medicine bottle, cannot follow a manual for performing a job. The ability to read basic material is one form of literacy, a prerequisite for many common tasks. Much more sophisticated language skills are indeed necessary for the increasingly technical nature of American society. In fact, the need for reading skills far beyond a minimal level is increasing as the number of unskilled jobs continues to decline. Hence, it is appropriate to examine how research can help society achieve increased levels of literacy in efficient ways.

One reflection of the importance of reading is found in the Adult Functional Reading Study, which attempted to assess the reading habits of a representative sample of 5,000 American adults (Murphy, 1973, 1975). The respondents reported that all aspects of their daily life involved some form of reading and that it involved a relatively large amount of time. In the study 71 percent reported reading for an hour or more per day. Reading is a particularly important skill for children, since it is a major cornerstone of our educational system. Not only is reading a school subject itself, but it is also a skill that is necessary to master other school subjects. It has been suggested that variations in reading skill may account for the high correlations among the grades a student receives for different school subjects, correlations that are typically .60 or larger (Bloom, 1976).

Another way to view the importance of reading skill is to consider the costs incurred by an individual who lacks reading skill. One cost, which is not easily measured, is incurred by society and by the individual when poor reading skills prevent him or her from obtaining a better job or from becoming a more informed citizen. A more easily measured cost is that of remediation, such as the large remedial programs in the armed services that are necessary because recruits cannot read the materials necessary to learn and execute military procedures.

This paper discusses current basic research on reading as a cognitive process, focusing on the mental processes that allow a reader to go from the printed page to a complex thought. The reason for emphasizing reading as a thinking process is that these processes can often be taught or improved. We present some of the major research findings on the acquisition and improvement of reading as well as theories about why some people are very poor readers while others are very good. We discuss research on the reading problems of special subpopulations and review some efforts to make written material more readable. In addition to considering what is currently known, we also suggest some interesting and important questions that should be explored in future research.

Reading and Understanding

Reading is a form of language understanding that shares many elements with the understanding of spoken language. In addition, of course, reading consists of some psychological processes that are specific to the visual processing of language. The contributions of the general comprehension processes to reading performance have generally been underestimated and sometimes entirely ignored. Reading comprehension has often been considered as something apart from listening comprehension. Consequently, reading problems are assumed to reflect something very specific about how a person extracts meaning from print. By contrast, if some reading problems are truly language problems, then they involve more general language skills. A related issue is the interchangeable usage of the terms _reading_ and _literacy_. Literacy sometimes involves much more than reading comprehension; when it does, it is important to be specific about its meaning and to distinguish it from reading. By reading we mean understanding the written equivalent of what could be understood through listening. A reasonable approach to the "reading problem" is one presented by Jenkins and Liberman (1972:1): "At all events the 'reading problem' as we know it would not exist if in dealing with language, all children could do as well by eye as they do by ear."

At a simple level, we can distinguish between reading problems and more general language or knowledge problems by testing whether someone can understand material when he listens to it. If he cannot, then the problem should not be construed as strictly a reading difficulty. For example, Sticht (1972, 1975) has found groups of young adults (military inductees) with very poor reading comprehension skills. Psychological tests showed that these people understood equally poorly when the material was presented orally. In this case, the problem should not be construed as a reading problem. It suggests that remediation may lie in teaching more general language skills along with other problem-solving skills and perhaps more specific knowledge relevant to the kinds of texts that will be read. At a simple, practical level, it makes sense to distinguish between general language skills, reading skills, and background knowledge. All are used in reading, but they point to different kinds of problems and different approaches to remediation.

HOW WELL DO AMERICANS READ?

Before we discuss reading research and its social implications, it may be useful to consider what a reasonable goal is for reading achievement for our society. In a report to the National Institute of Education, a group of prominent educators and scientists suggested that society's goal is twofold (Miller, 1973). One goal is a minimal level of reading (and writing) skill that is expected for the entire population. The ability to read common materials, such as newspapers, manuals, and instructions, ensures that people have sufficient reading skills to perform necessary, everyday tasks. The objective may seem limited, yet this criterion is stronger than that of any earlier period of history (Resnick and Resnick, 1977). In addition to basic skills, a second goal is allowing and encouraging people to increase their reading and writing skills beyond a minimal level (Miller, 1973). The opportunity to attain facility in reading, at least to the level of one's oral skills, enables people to contribute as much as possible to the intellectual, social, and economic activities of society.

While these are the current expectations of our society, it is interesting to note that the concept of literacy has changed over the last two centuries. It is only recently that society has expected fairly sophisticated reading skills from a large segment of the population (Resnick and Resnick, 1977). In the 19th century, mass literacy was not meant to include the entire society, and--even more strikingly--it did not entail general reading skill. For most "literate" adults, reading entailed only the ability to read aloud a limited number of very familiar religious texts. Only a very few adults were expected to be able to read other texts, to gain new information from them, and to read them critically. This latter group of skills is the current definition of the term reading.

It is commonly assumed that America has a reading problem, but estimates of the reading problem vary considerably. Precise demographic data are needed, however, to specify how well Americans of various ages and socioeconomic groups read and how well their skills match their needs. There is considerable evidence that high-level literacy skills have been declining over the last two decades and, given the increasingly technological nature of our society, this fact is cause for some alarm. Of even greater concern is the central impression that many Americans lack even basic reading skills.

Two assessment methods have been used to determine exactly how well Americans read. The most common is the norm-centered definition, in which a person's reading level is compared with some grade-level definition of reading. For example, a 9th-grade reading level is the level at which 50 percent of 9th graders pass some criterion score. Using this method, a national study found that 15-30 percent of 12th-grade students read below the 9th-grade level (Fisher, 1978). The U.S. Department of Defense found that 20 percent of 38,000 recruits read below the 7th-grade level (Sticht, 1979). This is of concern to the armed services because many jobs require basic reading skills. For example, army cooks have to read material at the 7-7.9th-grade level and supply specialists have to read material at the 12-12.9th-grade level (Sticht et al., 1972).

A second approach to measuring reading skill is the functional definition, in which reading is tested with material that is taken from real-life situations. One of the first such tests, carried out by Louis Harris and Associates (1970), attempted to assess skills such as completing the applications for social security benefits, a personal bank loan, public assistance, medicaid, and a driver's license. The study found that 8 percent of the representative sample of 1,685 people answered less than 80 percent of the questions. A more recent study by the National Assessment of Educational Progress (1976) examined the skills of 17-year-olds using items that were relatively easy everyday questions. The study found that 12.6 percent of the students answered fewer than 75 percent of the questions correctly; those scoring less than 75 percent were deemed illiterate.

Recently, Fisher (1978) has argued that the classic studies of functional literacy may overestimate the size of the literacy problem among high school graduates. He argues that some questions tap more than basic reading skill and are often unanswerable by managers and professionals. Correcting for this and for a 1 percent failure rate due to boredom or fatigue, Fisher argued that the surveys can be reinterpreted to show that as few as 0.6 percent and certainly no more than 7 percent of high school graduates are functionally illiterate. Moreover, since these literacy tests tap skills and knowledge that are not specific to reading, an error could reflect problems with language understanding, background knowledge, or reasoning--not necessarily reading. There is some disagreement over the extent of the literacy problem for high school graduates. There is little disagreement, however, about the problems of people who repeat grades and drop out of school prior to high school graduation. Dropouts consistently have lower achievement scores than students who stay in school. These people are not included in the literacy rates gathered on high school seniors. Including dropouts in the National Assessment of Educational Progress literacy sample estimated in 1975 may yield estimates of illiteracy as high as 20 percent of 17-year-olds (Lerner, 1981).

As we mentioned earlier, functional reading skill is a minimal goal for society. Much more than functional skill is necessary in a society that is becoming increasingly technological and in which higher levels of education are necessary to take advantage of the major employment opportunities. Correspondingly, it seems necessary not only to achieve but also to maximize reading skills. Thus, basic research in reading should not be aimed exclusively at understanding the causes and remediation of extremely poor skill. It should also be aimed at understanding skilled reading and how it can be achieved. In fact, much of the basic research in reading to this point may contribute toward this goal, by analyzing the components of higher levels of reading skill in order to improve reading instruction.

CURRENT RESEARCH IN READING

There is something new in the field of reading. As the result of developments in the disciplines of psychology, linguistics, and computer science, empirical and theoretical advances in the study of reading have been fairly dramatic. These advances tell us what a reader's eyes and brain do during reading that enables him or her to comprehend written language. Our knowledge of the reading process is so detailed that we can make a reasonable estimate of how long it will take (down to fractions of a second) for a college-age reader to comprehend a given word of a passage as well as an estimate of what he or she will understand and recall from the passage. We have progressed beyond predicting reading difficulty on an actuarial basis to a better scientific understanding of the content and outcome of the reading process.

The core intellectual basis for the study of reading is in the discipline of psychology. Psychology attempts to specify the mental processes that transform a series of words printed on a page into a coherent thought in the mind of the reader. In the early part of the 20th century, researchers began an interesting analysis of reading and its acquisition. Although early researchers lacked current theoretical and methodological tools, their observations are still of interest. Like most sciences, psychology undergoes periodic changes in its theoretical outlook, and changes in psychology in the 20th century have had a large influence on the study of reading. Between approximately 1935 and 1955, the dominant philosophy of behaviorism proscribed the study of complex cognitive processes such as reading and dampened research progress in this area.

The next paradigm, the information-processing revolution, was entirely congenial with the scientific study of reading. Beginning about 1955, information-processing psychology arose from developments in information theory (a branch of applied mathematics), symbolic processing by digital computers, and the psychological study of complex human behavior. Information-processing psychology focused on precisely those internal processes, in their full detail, that behaviorism did not acknowledge. In this approach thought consists of information processes, operations whose operands and outputs are mental symbols. Mental symbols represent knowledge acquired from the outside world and from internal computations. From this perspective the questions of interest about any behavior are:

1. In performing this task, what information is internally represented, in what form is it represented, and what new (partial and final) information is acquired in the course of performing it?
2. What processes are used in this task, on what input information do they operate, what do they output, under what conditions are they invoked, how long do their computations take, and what are the sources of errors?
3. How are the processes acquired?

While these questions have been asked, and successfully answered, about many tasks, only recently have they been applied to reading.

An important contribution of the information processing approach is its acknowledgement of the role of the environment in explaining human behavior. The direction and content of thought processes can be shaped as much by current circumstances as by previously acquired knowledge and thought patterns. Strategic behavior is exquisitely adaptive, so that variations in any laboratory task evoke corresponding changes in a person's strategies. As applied to reading, this means that how a person reads and what he or she understands and remembers depends not only on knowledge and reading ability but also on the text that he or she is reading and on the particular reading situation. People read differently depending on whether they think they are going to be tested on the material, and even depending on the particular type of test they anticipate. One straightforward demonstration indicated that two groups of subjects recalled a description of a house very differently depending on whether they read it from the perspective of a prospective purchaser or a burglar (Pichert and Anderson, 1977). It became clear that theories of reading would have to be accompanied with complete analyses of the situations in which reading occurs.

The information-processing approach analyzes reading in terms of its component processes, decomposing the complex whole into more understandable parts, such as word decoding, syntactic analysis, semantic analysis, and analysis of the situation referred to by the text. This decomposition has obvious benefits, although a less obvious drawback is that the parts are not equally understandable. Even those that are understandable are not easily reconstituted into a theory of the entire reading process.

The problem of how various subprocesses in a complex task work together was a difficult one for psychologists, requiring some expression of how a complex multicomponent system may function in a coordinated, goal-directed way. The solution came from the field of computer science, which provided an example and a medium for constructing large-scale theories of complex behaviors composed of elementary, understandable parts. When expressed as a computer simulation model, each component process is specified in detail, and the many component processes function in coordination to perform some complex task.

Artificial Intelligence

The branch of computer science that deals with artificial intelligence attempts to develop computer programs that intelligently perform significant tasks. Unlike many other branches of computer science, artificial intelligence makes use of symbolic but nonnumerical computation. Artificial intelligence programs make use of knowledge, induction, deduction, and perception in order to make the computer program behave intelligently. One of the first symbolic-processing programs, written in the late 1950s by Newell, Shaw, and Simon (1957), was able to prove many of the formal theorems of logic from Whitehead and Russell's Principia Mathematica, a dramatic demonstration that computer programs could simulate what we consider to be human thinking. In the 20 years since then, striking technological and scientific advances have produced programs that play chess and understand language within specified constraints. The technological advances include programming languages that are expressly designed for symbolic (as opposed to numerical) processing, programming languages that are designed to mimic the organization of human thought, and hardware advances involving computers with increased storage capacities and speeds at decreasing costs. This progress in computer science has provided psychological theorists with a medium to express theories accounting for complex behaviors. The expression of a theory as a computer program imposes a rigor that forces every functioning component of the theory to be specified in detail in a formal language that is unambiguous. Furthermore, the successful execution of such a program provides a demonstration that the theory is logically sufficient to account for the behavior it is performing.

There are a number of computer simulation models of natural language understanding that provide some insight into human reading. These programs are invariably large and complex but share certain properties that are also apt to be shared with human readers. One commonality is that the programs must have a great deal of knowledge about the words and structure of the language in order to comprehend successfully. Many of the programs manage to have vocabularies of only a few hundred words, and few vocabularies are more than 1,000 words. The word knowledge necessary for comprehension appears to increase exponentially with vocabulary size, since for each new vocabulary entry information must be added indicating the relationship of the new entry to all the relevant previous entries. The amount of information that must be added increases with the number of entries.

Successful programs also have a considerable knowledge of the topic of the text. This was first made clear in Winograd's (1972) program, which understands questions and instructions about a restricted world made of toy blocks. The lesson was even more forcefully brought home when programs turned to more naturalistic texts. Charniak (1972) illustrated the point by detailing the specific knowledge that the reader must provide to understand a simple child's story describing a little girl going to a piggy bank after hearing the ice cream truck bell. Knowledge has proven to be crucially important to systems that understand human speech. A speech understanding system can recognize human descriptions of chess moves if the program knows what moves are legal and likely and if the program itself can play chess (Reddy, 1980). The moral that might be taken from this result to the field of reading is that good comprehension depends on familiarity with the subject matter. Later we present psychological evidence that this moral applies not only to computer simulations but also to human readers.

A third commonality is that all successful computer understanding programs analyze several aspects of the text, including its lexicon, syntax, semantics, pragmatics, and morphology. No program that relies exclusively on syntax or exclusively on semantics is successful; different kinds of analyses and knowledge must be brought to bear to achieve comprehension. Moreover, the use of these various levels of analysis must be effectively coordinated, so that they can collaborate on comprehending a given piece of text. The lesson here is that training should aim at many levels, such as vocabulary, grammar, and text composition. It is unlikely that an exclusive focus on one level will be sufficient.

Linguistics

In the late 1950s, the field of linguistics (and several related disciplines) was jolted by Chomsky's (1957) study of English syntax. While Chomsky's particular theory is probably of less direct relevance to psychological models than was originally thought, his work did contribute to the current basic research in reading in two important ways. First, Chomsky's work and subsequent linguistic analyses have shown the importance of formal models of language. In brief, researchers now pay as careful attention to the structure of language as they had in the past paid to the nature of their population of readers. As well as providing tools for describing some aspects of language, linguistics has pointed out the importance of how the language of a text is structured. In addition to theoretical models, this general attention to the structure of language has resulted in increased sensitivity in several areas related to reading. For example, there is now renewed attention given to the structure of texts given to young readers, to the relationship between written and oral languages, and to how dialects may differ from the language expressed in a text. Part of this attention has arisen from the emphasis in linguistics on the nature of language.

Second, Chomsky shifted the focus of language research to the sentence from lower units such as words and morphemes by initiating the large-scale study of syntax (grammar) as opposed to the study of phonology. This level of analysis of language is more compatible with the interests of psychology and pedagogy, so many research and instructional programs were developed that revolve around the syntactic analysis of language. Since then the emphasis has moved to semantics and to units of text analysis that are larger than the sentence. In the 1970s, linguistic analyses were applied to the structure and content of extended texts, from paragraphs to short stories. The structure is characterized both in terms of low-level propositions (units approximately the size of a simple clause) and in terms of higher-order abstractions, sometimes referred to as micro structure and macro structure, respectively (Kintsch and van Dijk, 1978). Story grammars were developed to characterize the structure of narratives in terms of the components of setting, conflict, and resolution, and many studies have shown that people use such components in organizing the information they read in a story (e.g., Mandler and Johnson, 1977; Rumelhart, 1975; Stein and Trabasso, 1981). Schema theories characterize the organization of knowledge of conventional occurrences and its use in text comprehension. For example, Schank and Abelson (1977) propose a schema for organizing knowledge about what occurs in restaurants. Possible occurrences described in a text could deal with entering, ordering, eating, paying, and exiting, and each of these in turn can be further decomposed. Eating, for example, can be segmented into the cooking, the delivery of food to a waiter, the waiter's delivery of it to a customer, and the customer's ingesting it. There has been much discussion and some research on knowledge-driven comprehension, which appears to be an important factor in explaining how we read texts that deal with very familiar subject matter. The general suggestion is that knowledge of the subject matter tells the reader how to interpret the text in a meaningful way.

Psychology

Reading research in psychology attempts to explain how the printed symbols on a page are translated into a meaningful representation. To make this translation, the reader uses not only the information in the text but also his or her own knowledge about the topic. In keeping with the definition of reading as the processing of language by eye, it is possible to look at the research on reading from two perspectives. In some respects the research reveals fundamental aspects of language processing that are general to both reading and listening comprehension. In other respects the research focuses on aspects that are specific to the visual processing of written language.

In our analysis as well as in our discussion of social implications, we distinguish between skilled reading, unskilled reading, and beginning reading. The reason for studying skilled reading, aside from intrinsic scientific interest, is to understand how normal mechanisms work in order to understand deviations from normalcy. The analysis of the entire spectrum of reading abilities, from superior to typical to beginning to poor, constitutes good science and may also suggest ways of preventing or remediating reading difficulties. Reading involves processes and structures that are learned and can be modified. An understanding of skilled reading should indicate the end point of the learning process and provide some clues to how to attain it.

Reading includes a large number of cognitive processes, from the registration of the visual print to the final complex thoughts that may be triggered by the text. One of the goals of current research is to understand the reading process by examining each of the components in finer detail.

Decoding

The first step in reading is to register the printed text. This process of decoding has received a great deal of attention from researchers and probably constitutes the bulk of the research related to reading (see Gibson and Levin, 1975). There have been three central issues raised in this research. One is whether a reader must use the printed word to retrieve the sound and then use the sound to retrieve the meaning of a word. The advantage of such phonological mediation is that the processes used in speech understanding could then be used directly in reading. Many readers seem to hear the words as they read introspectively, suggesting that there may be some phonological component in reading. Most of the evidence suggests, however, that skilled readers generally do not use a phonological code (Bradshaw, 1975). Some simple evidence is the difficulty we have in reading phonologically correct but orthographically anomalous sentences: "Eye Do Knot No What You Herd" (Baron, 1977). However, young children or adults dealing with difficult words may rely on a mediating phonological code. If the material is very difficult or if the reader is not very skilled, he or she may even make lip movements (Hardyck and Petrinovitch, 1970). Some popular reading improvement courses try to help readers become more skillful by suppressing these minimal movements when the material is easy. In sum, it appears that part of skilled reading is the ability to proceed directly from the print to some representation of the word. The skilled reader uses an intervening phonological code only if the word is unfamiliar or if the material is particularly difficult.

A second major issue in this area is the unit of word decoding--the letter, letter clusters (such as ch), syllables, or the whole word. One very simple model of word perception holds that the reader identifies each letter and combines them to identify the resulting word. Such a simple model has difficulty in coping with the very pervasive finding that it is easier to identify a letter that is embedded in a word (like identifying the letter a in cat) than when it is embedded in a nonword (as in tac) or presented alone (Baron and Thurston, 1973; Wheeler, 1970): This result is called the word superiority effect. The evidence suggests that word recognition does not depend exclusively on serially identifying individual letters. Rather, the reader works on identifying several letters at once, and information about one letter helps in identifying other letters and vice versa (McClelland and Rumelhart, 1981).

A third issue in this research is that words themselves are easier to identify if preceded by a semantically related context. It is easier, for example, to process the word doctor if one has very recently processed the word nurse (Meyer and Schvaneveldt, 1971). The general explanation for this kind of semantic facilitation of word recognition is that accessing a given concept also activates some semantically related concepts. The prior accessing of nurse activates semantic relatives of nurse, among them doctor, and these activated concepts are then easier to retrieve. The results of such laboratory studies suggest that a rich knowledge of relationships among words may facilitate decoding and perhaps other levels of processing in the course of reading.

While we assume that all adults are skilled at word recognition, recent research suggests that there may be individual differences among readers. Good readers may recognize and retrieve words slightly faster than poorer readers (Hunt et al., 1975; Jackson and McClelland, 1979; Perfetti and Lesgold, 1977). The suggestion is that if it takes longer to recognize and retrieve a word, other processes that depend on this retrieval will have to wait longer and therefore be subject to forgetting. This correlation accounts for a relatively small number of the differences in the reading comprehension of adults, but it is a consistent finding and seems to be part of what makes some readers better than others.

Learning to Read

While phonological decoding may not be necessary in adult reading, it is critical in learning to read. Young readers are expected to learn how to decode visual symbols into sound. Children who cannot do this are the "poor readers" (Rozin and Gleitman, 1977). But this may be true only in the early years, when children are learning to read. After about 3rd and 4th grade, children have sufficiently mastered decoding skills, and their individual differences do not seem to closely reflect word decoding abilities (Curtis, 1981).

Since word decoding is crucial to early reading, it is useful to examine research that suggests the necessary prerequisites for cracking the alphabetic code. Evidence suggests that awareness of the phonological structure of language is one predictor of reading success. For example, five-year-old prereaders who can segment a spoken word into its constituent phonemes (e.g., say _table_ without the _t_) will tend to be better in a word recognition test at the end of 2nd grade (Liberman et al., 1977).

Another research program on prerequisite word decoding skills has taken advantage of the fact, now well supported, that children find the concept of the phoneme (like the _b_ in _bit_) a difficult one. But they find it relatively easy to abstract the concept of a syllable. Rozin and Gleitman (1977) constructed a reading curriculum that segmented words into syllables and provided pictoral hints to the meaning and pronunciation of the syllables. For example, the word _inside_ was divided into two parts, _in-side_, and pictures, one associated with _in_ and another with _side_, were presented above the appropriate syllable. The curriculum was used as the sole initial reading program in seven 1st-grade classrooms from three inner-city Philadelphia schools, in which the reading achievement norms were either average or considerably below average. Children with a poor prognosis for reading acquisition did not read well during the 1st grade. They did make substantial progress, however, in acquiring some of the components of reading skill. The researchers believe that the curriculum had a substantial motivational effect (Rozin and Gleitman, 1977). Other research on the teaching of phonological concepts and segmentation is appearing in other curricula (Resnick and Weaver, 1979) and seems to have had some success. One important feature of this approach is its combination of field research and theoretical analysis.

Knowledge and Comprehension

When we read (or listen), we often rely on previous knowledge to guide comprehension. At the extreme, it has been shown that if we have no clue to the topic of some passage, then it is very difficult to understand it, much less remember it (Bransford and Johnson, 1973). People who know something about a topic understand new information about that topic more easily and remember it better (Spilich et al., 1979). For example, a baseball expert can understand and recall a passage about a baseball game much better than a baseball neophyte.

Knowledge may affect comprehension in several ways. It may provide the vocabulary the reader or listener needs to read the passage, yet its influence may be at once more subtle and more dramatic. Knowledge is internally organized and its organization may provide a preexisting framework that the reader can use for assimilating the new information. This means that the reader already knows something about probable relationships, if not specific connections, described in the text. She or he already has some idea about what is important and what is not, about what is likely to happen and what is not. Knowledge provides a structure within which the new information can be framed, saving the reader from having to infer both the structure and the specific new ideas.

Knowledge of particular kinds of text structure, such as narrative, has recently received a great deal of attention (Stein and Trabasso, 1981). Narratives often begin with a setting, introduce a character who has some goal, and involve some complication that must be resolved through actions. Even young children are aware of this structure and use their knowledge of it in understanding stories that are read to them. Recent analyses of grade school reading materials suggest that many selections violate normative story structures, making it difficult to extract the coherence of the story (Anderson et al., 1980; Tierney et al., 1980). It may be better to give young children stories that are more predictable and that follow conventional structures, in part because it gives them an opportunity to acquire and make use of such structural knowledge.

Mental Chronometry

Another important development in the field of psychology was the advent of mental chronometry, the measurement of the duration of mental events. This was brought about in part by the development of new experimental paradigms and associated theory to examine the time course of various mental processes. The chronometric approach tends to focus on low-level processes (lasting from fractions of a second to several seconds) and simple tasks. For example, many studies investigated how the contents of short-term memory was scanned for the presence of a given item (Sternberg, 1969). The availability of laboratory computers in the 1960s made this kind of research easier to do. The information-processing theories that evolved from a chronometric approach attempt to characterize the flow of information through the mind, with particular emphasis on the relative time course of various processes involved in the flow.

Eye Fixation Research

A further evolution of mental chronometry came with the use of eye-tracking to measure the duration and sequence of mental processes. While ordinary chronometric research simply measures how long it takes to perceive a single item or to perform a simple task, eye fixation research measures how long and in what sequence a subject looks at each item in a display in trying to solve a problem. The problem in the study of reading is to comprehend a passage of text. This approach provides a detailed characterization of where and for how long a reader looks in reading a text, where he or she pauses, what he or she skips, what he or she looks back to. The analysis breaks down the behavior into fractions of a second, with the average looking time at a word being about a fifth or a quarter of a second (approximately 300 words per minute). This kind of research was impossible on a useful scale before the availability of laboratory computers to determine where the reader's eye is pointed and to record the acquired data.

Eye fixation research done in the past decade has revealed a number of important properties of reading (Just and Carpenter, 1980; Rayner, 1975). Some of the recent results answer questions that were posed at the beginning of the century about the perceptual aspect of reading. First, the perceptual span in reading is fairly small, approximately two or three words wide. That is, when a reader's eyes are pointed at the nth word of a passage, he or she may possibly be able to read the word before or the word after without moving his or her eyes, but beyond that he or she cannot read the words (although he or she may be able to determine something about the length and shape of the words). Thus the perceptual window through which we view a text is fairly small. In fact, readers fixate on most content words of a text that is technical or unfamiliar, between 70-85 percent. They fixate on only about 40 percent of the short function words, such as a or of. The idea that a reader can make only one or two fixations on a line or on a page during normal reading is a myth.

An important theoretical step has been to use the characteristics of eye fixations and relate them to the detailed mental processes that have been postulated in psychological and artificial intelligence models. Thus, if a word initiates a difficult process, the reader would be expected to spend more time on that word. Just and Carpenter (1980) found such a result. The durations of the pauses that readers' eyes make on various words of a text are determined by the characteristics of each word and the moment-to-moment processing to which it is subjected in a given linguistic context and task context. The ongoing psychological processes control the duration of the pause on each content word. For example, less familiar words are looked at longer (up to a fifth of a second longer) than familiar words. The significance of this result is that one can determine the processing load or difficulty at each point in a text by measuring the duration of the pause. A second result from this research is the finding that words are generally given an interpretation as soon as the eye encounters them. Readers attempt to understand everything as soon as they see it, rather than postponing the choice of interpretation until more data (from the remaining parts of the sentence) are collected. Sometimes it is necessary to postpone the interpretation, but it is usually made immediately. If the immediate interpretion later proves to be incorrect, then efficient correction procedures pinpoint the source of the error; readers quickly and accurately look back to the word or phrase that is the source of the problem and choose an alternative interpretation that is consistent with the text that follows. This is a model based on experience with skilled readers that provides a baseline for comparison with less skilled readers, including those with general language deficits or those with specific reading disabilities.

This detailed chronometric analysis meshes very well with the linguistic analysis of texts, so that many factors coalesced to produce rather complete data-based theories of reading. It became possible to determine the amount of processing time spent on each part of a text (at the level of words, phrases, or sentences) and relate that performance to a linguistic analysis of the text structure. The proposed analysis can be expressed as a theory of information processing, written as a computer program that itself reads text and has similar performance characteristics to human readers. It can pause when it has difficulty in comprehension and speed up when the comprehension is easy, just as human readers do. It can produce a summary of the text it has just read or answer questions about it, about as well as human readers do. This is the current state of the art in the field of research on reading.

COGNITIVE DEVELOPMENT AND READING ACQUISITION

Teaching Methods

The history of reading instruction can be viewed as a tug-of-war between two main instructional methods: the whole-word method and the phonics method. Phonics was the method of choice at the turn of the century, with an emphasis on drills and memorization by rote long before the child was introduced to stories. In reaction to this, the whole-word approach arose in the 1920s, with an emphasis on sight recognition of words, and it prevailed virtually unchallenged between 1930 and 1950. In the 1950s and 1960s, there were two challenges to the whole-word method (Beck, 1981). One was Flesch's (1955) book, _Why Johnny Can't Read and What You Can Do About It_, a denunciation of teaching practices in reading that received considerable popular attention and acclaim. A careful evaluation of the evidence did not appear, however, until Chall's (1967) book, _Learning to Read: The Great Debate_, in which she carefully analyzed existing programs and their effects on reading. She concluded that there was a difference between the two methods and that the phonics approach appeared to have the advantage, at least in the grades in which the comparison was made, grades 1-3. The current scientific consensus seems to tilt in favor of phonics. There is no evidence that the phonics approach impairs comprehension and, since the child must master sound-symbol correspondences at some time, it makes sense to teach the correspondence directly and to do so from the beginning.

Currently, it is argued that most reading programs combine the two approaches to reading instruction. However, in a recent analysis of eight major reading programs, Beck (1981) has argued that the distinction still exists. She analyzed eight major early reading programs and found two distinct types.

Programs that have traditionally been associated with whole-word approaches continue to reflect that approach, even though they include some phonics. These programs, which she termed basal programs, can be distinguished on a number of dimensions from those that stress a phonics approach. The basal approach introduces the new words of a story early in an instructional episode and, if it deals with relevant phonics principles, it does so later in the episode. The newly introduced words in a lesson seem to be frequently used words, rather than ones that have regular grapheme-phoneme correspondence. Thus, if the child is trying to learn the sound-symbol correspondences, he or she must work with words that are not the best examples in a basal curriculum. In addition, the basal programs use a large proportion of words that cannot be phonetically decomposed using the phonics rules introduced to that point. Finally, the basal approach gives little practice in blending, that is, combining individual sounds into a single word. Such a process is considered extremely important in reading acquisition.

For the programs using a phonics approach, almost all the words in the stories conform to the rules introduced to that point. The phonemes are named or isolated in contrasting words (e.g., pat versus bat). In sum, there still are significant differences in various programs, but their subsequent impact on the development of skilled reading remains unassessed.

Basic research in reading has also suggested that fluency in word recognition is an important component of skilled reading. It has been shown that if word recognition is slow, it can interfere with subsequent comprehension and that word recognition time is consistently correlated with reading skill. It would therefore seem advantageous to promote practice in word decoding, to ensure that children achieve fluency. Textbooks, however, show a trend toward larger vocabularies and less repetition. Rodenborn and Washburn (1974) reported that in the older basal programs "a word was repeated from 6 to 10 times on the pages immediately following introduction" (p. 886). In current basal programs, they found, the majority of words occurred fewer than 6 to 10 times. Beck (1981) notes quite reasonably that it may not be a good practice to move children toward increasingly difficult material, irrespective of the fluency with which they handle current material. Further research may be necessary to examine how fluency affects the mastery of subsequent levels.

Another aspect of reading that has been stressed in current research is its interaction with the reader's knowledge. The information from the text must be related to what the child knows. We know that knowledge is one important component of what develops in a child, but psychology is only starting to examine how the knowledge acquisition process may interact with specific skills such as reading.

Metalinguistic Skills

Reading and listening comprehension are both rooted in more general cognitive skills that seem to be acquired developmentally. For example, consider the very basic skill of knowing when you don't understand something and what it is that you don't understand. Children are not very sophisticated in these skills. For example, they do not detect obvious inadequacies in instructions, they may accept contradictory information, and they often fail to ask for clarifying information when it is needed (Brown, 1980; Markman, 1977). Adults show more sophistication. For example, adults tend to spend more time on important ideas when reading a passage (Just and Carpenter, 1980), perhaps in part because they can distinguish essential from inessential information. By contrast, children are much less successful in distinguishing the relative importance of information (Brown and Smiley, 1977). This has clear implications for how they might try to study a text. Because older children and adults do better on distinguishing among the kinds of information in a story, they can more effectively use prolonged study time, whereas children below 7th grade do not (Brown, 1980). The analysis of the problem-solving strategies relevant to reading and listening comprehension are only beginning to receive attention.

Children begin formal reading instruction in 1st grade, usually at the age of six. Psycholinguistic research has focused primarily on the linguistic competence of much younger children, those between ages one and three. The implicit assumption seems to be that children have acquired "language" by the time they are three or four years old, or at least the interesting syntactic and conceptual basis of it, even if vocabulary development continues into adulthood. The other implicit assumption seems to be that linguistic competence is equivalent across children. It has recently become clear, however, that neither assumption is warranted. Children continue to show linguistic development considerably beyond the early years, and they vary in spoken language competence. Again, the precise relationship between stages of spoken language competence and reading acquisition and skill has generally not been explored. With increasing emphasis on language development beyond the early years, it may be expected that its relationship to reading will soon receive more attention.

Finally, cognitive psychology has recently turned its attention to more general issues of skill acquisition and its relationship to cognitive development. Such research may also help in understanding the prerequisites of attaining high levels of skill in any cognitive task.

SPECIFIC READING DISABILITY

There is a population of children who have inordinate difficulty in learning to read, compared with their performance in other areas. These children are classified as dyslexic (particularly if they are being studied by medical researchers) or as having a specific reading disability, to distinguish their problem from children with general intellectual difficulties. We will use the term specific reading disability, although the term is something of a misnomer, because the disability is often not specific to reading but may reflect a more general language comprehension problem. Specific reading disability has attracted some research attention from educational, medical, and psychological circles (see Benton and Pearl, 1978). However, it appears to not have benefited from recent cognitive research.

One major difficulty with research in this area is that there is no general consensus about the nature or definition of dyslexia (Rutter, 1978). For research purposes, children are selected by excluding those who have "good reasons" for their reading problems. For example, children who have general intellectual deficits, obvious vision or hearing problems, or marked emotional problems are not selected. If a child does not fall into these categories but still lags two years or more behind peers in reading, he or she may be classified as having a specific reading disability. However, the definition by exclusion does not necessarily result in a homogeneous population of children.

It is increasingly recognized that some dyslexic children may have difficulty not only with reading but also with some basic language functions. Initially a rather popular hypothesis was that reading disability was visual, and there were reports of confusions between letters such as *b*, *d*, *p*, and *q* or between the words *saw* and *was*. Such reversals tend to be infrequent. One study accounted for only 25 percent of the errors in lists that were constructed to maximally allow for such reversals, suggesting that these errors are not a major source of the difficulty in dyslexia (Liberman et al., 1971). In addition, poor readers "see" such letters and words in an unreversed order, as evidenced when they are asked to copy them (Vellutino et al., 1975). That the children may reverse the words in naming but not in copying suggests that their limitation does not lie in the graphic encoding. Similarly, normal and dyslexic readers may not differ with respect to reproducing geometric designs (Vellutino et al., 1975). Several results suggest that the orientation and sequencing inaccuracies observed in the reading and writing performance of some poor readers may reflect linguistic problems rather than perceptual distortions (Liberman et al., 1971; Vellutino, 1977).

At one point a theory was suggested that dyslexic readers have particular difficulty in associating visual and verbal elements (Birch, 1962). Since the original formulation, however, it has been shown that poor readers have a great deal of difficulty in verbal coding and that such problems could explain problems with visual-verbal association (Bryant, 1968; Vellutino, 1977).

Another hypothesis that seems unlikely is that poor readers suffer from an inability to maintain information about sequences. But there is currently little support for the hypothesis that poor readers are poor only in extracting sequence information and not in extracting item information. The suggestion is rather that poor readers have difficulty with verbal coding (Vellutino, 1977).

More recently, the hypothesis has been put forward that poor readers have difficulty with some specifically auditory-linguistic aspects of either reading or learning to read. For example, it has been reported that a large number of children referred to clinics for reading problems have a history of speech and language disorders (Ingram and Reid, 1956; Ingram et al., 1970; Lyle, 1970). Such reports have the problems associated with being retrospective and in being somewhat uneven as to the population being studied (Vellutino, 1977) but are nevertheless suggestive. They indicate that poor readers generally have smaller speaking vocabularies, less verbal fluency, and sometimes use grammar and syntax inappropriately (Fry, 1967; Schulte, 1967).

Several studies have suggested that dyslexics have difficulty in retrieving words. For example, Denckla and Rudel (1976a,b) found that they take longer than normal readers to name common objects, colors, words, and letters. Poor (but not necessarily dyslexic) readers have been found to take longer on other word-naming tasks (Perfetti and Hogaboam, 1975).

Researchers cannot assume that specific reading disability refers to a single characteristic. Some evidence suggests that there are several distinct kinds of reading disability. There is also a confusion of cause and effect. Researchers may be studying an effect or a corollary of reading dysfunction, rather than a cause. One possible research policy on specific reading disability is to encourage more joint efforts by psychologists, medical researchers, and educators to examine individuals in detail as well as to encourage more longitudinal research.

Effect of Dialects on Learning to Read

There are many ways in which the written language does not exactly match spoken language. These differences and their possible effects on reading are only now being examined. For example, in speech we have gesture and intonation to disambiguate the major focus of a sentence. Moreover, the feedback between the speaker and listener and the general conversational spirit of most speech has no counterpart in reading. There are also some syntactic differences, such that some grammatical structures are used much less frequently in speech than in writing. For example, cleft constructions like "It was John who won the trophy" seldom occur in spoken language. But they are useful in writing because they allow the writer to stress that the new information is the identity of the winner and presuppose that the reader knows that someone won the trophy. In speech, much the same effect can be accomplished by vocally stressing the name John, "John won the trophy." Because of these differences between spoken and written language, the successful reader must be prepared to accept a different linguistic style in reading. These differences may seem small, but for some English dialects they are considerable. In particular, it has been suggested that black English constitutes a dialect that is sufficiently different that there may be interference between spoken and written language (Hall and Freedle, 1975; Stewart, 1969).

Research has begun to document and analyze the differences between black English and standard English. One level consists of phonological differences. For example, black English has vowel variations (as do other dialects, such as the Bronx or a Southern dialect) as well as other phonological differences, such as weakening of the final consonant. The evidence on whether such phonological differences are a source of reading difficulty, however, is conflicting and weak (Hall and Guthrie, 1980).

There are also some syntactic differences between black and standard English. For example, "He be busy" in black English means "He is busy" in standard English because be is a marker for habitual action. In black English, the -ed suffix is typically omitted from spoken language; Labov (1970) found that this may cause miscomprehension of the tense of verbs. The evidence that these differences cause reading problems is inconclusive. One reason for the absence of strong evidence may be that black children acquire standard English as a second dialect. Thus, the effect of dialect differences may decrease with age. One of the most striking effects of dialect on story comprehension was observed with 4 1/2-year-old children (Hall et al., 1975). It was found that black children remembered stories in black English as well as white children remembered stories presented in standard English, and both did worse at remembering stories in "foreign" dialects. Hall and Guthrie (1980) suggest that the effect of dialect differences may be construed more broadly in terms of cultural differences. They suggest that research could examine differences in the patterns of language involved in nonverbal as well as verbal communication.

Reading by Deaf Children

The intimate relationship between language and reading manifests itself in the problems that deaf children have in learning to read. The reading level of deaf children is markedly below that of their normally hearing peers. Relatively comprehensive studies in the United States (DiFrancesca, 1972) and in Great Britain (Conrad, 1977) have found that deaf children between the ages of 15 and 16 read at approximately the level of hearing 9-year-olds. The extent of reading impairment is directly related to the amount of hearing impairment. The profoundly deaf (often defined as those who suffer a hearing loss greater than 85 dB) read considerably more poorly than those who have a less severe hearing loss. The least deaf 15-16-year-olds in Conrad's study in Great Britain read about as well as average hearing children at age 10.5, while the most deaf had the same reading ability as average hearing children at age 8.2. The implication is that deaf children often have difficulty acquiring functional literacy because of their hearing impairment.

By one recent census count, there are about 450,000 profoundly deaf people in the United States (Bellugi et al., 1974-1975). The number who are prelingually deaf (whose deafness occurred before they acquired a spoken language) is much smaller. Also, whether a deaf child is born to deaf or to hearing parents affects later language performance. Deaf children born to deaf parents tend to acquire American sign language (ASL) as a natural primary language. Such children constitute about 10-15 percent of deaf people. The advantage for these children is that they learn ASL from early childhood, the way a hearing child learns English. Children of hearing parents have difficulty acquiring oral language and they often acquire sign language, although they will tend to learn it later, from deaf peers and in schools that teach sign communication. While it is possible to sign English, ASL is the language most often used by the deaf among themselves and in deaf families.

ASL is a language that is conceptually, morphologically, and syntactically distinct from English. Recent research has only begun to explore the structure of this language and its acquisition (Klima and Bellugi, 1979). ASL has several properties that reflect its character as a visually communicated language rather than an oral language. In English, for example, there are phonetic segments, such as p, ae, and s, that can be combined in different orders to create different words, such as pass, asp, and sap. Hearing subjects who are given a written list of words and asked to recall them may make sound-based errors, such as recalling vote as boat, and this is taken to indicate that they stored the material in an acoustic form. ASL does not have phonological segments; its structure is based on gestural units, such as hand configuration, the place of signing, and the type of movement. These are also the features that are used in the teaching and learning of ASL. Deaf subjects fluent in ASL who are shown a series of signed words will subsequently make gestural errors during reading, such as confusing the direction of motion or the place of signing (Bellugi et al., 1974-1975). This suggests that the visual form of the language determines its coding in immediate memory.

ASL differs from English not only in its modality (gesture versus speech) but also in its lexical structure. Just as it is often difficult to make exact translations between English and French, so is it difficult to make exact translations between English and ASL. The languages differ in syntactic structure as well. The fact that ASL is such a different language makes the acquisition of written English fairly complex. A deaf child learning ASL is gradually gaining linguistic competence in one language (ASL) and learning to read in another (English). Learning to read a second language without speaking it is not uncommon for literate adults, but it is very unusual for children learning reading for the first time. Deaf children reading at the level of a 10-year-old do not supply a word missing from the text the way a hearing 10-year-olds do (Moores, 1971). Some of the difficulty in learning to read may arise from the differences between English and ASL; however, a simple difference explanation does not seem to adequately account for the data. There is evidence, for example, that deaf children who acquire ASL from their parents, and presumably gain fluency at an earlier age, understand ASL better and also understand manually coded English better (using ASL with English word order, functor words, and important affixes either signed or finger spelled) (Hatfield et al., 1978). There has been some suggestion that children who are more fluent signers, because they acquired the skill as young children from deaf parents, are better readers (Stuckless and Birch, 1966). It has been suggested that one important problem with current approaches to education for the hearing-impaired is that their exposure to language instruction too often begins after the optimal period for language learning.

The reasons for reading problems with deaf children are complex. It is possible that deaf children have less rich experiences to draw on than hearing children and that this deficit is revealed in reading comprehension. Such a deficit could also reflect differences in educational treatment and effectiveness. In this regard, it is interesting that the results for the United States and Great Britain look remarkably similar. While precise data are not available, there have been reports of difficulty in teaching the deaf to read in many countries, such as the U.S.S.R. (Moores, 1972), Sweden (Ahlstrom, 1971), and Denmark (Vestberg, 1973). Reading acquisition by deaf children is likely to benefit from further understanding of the relationships among spoken, signed, and written language processing.

BETTER WRITING

We have discussed reading problems by focusing on the reader. In some cases, however, the problem is in the text. When the goal is efficient and accurate communication, it may be more useful to improve the text. It is generally agreed that many forms, such as insurance policies, home mortgage applications, and leases, are often incomprehensible. The people who are supposed to read and benefit from a particular government program often read at a level far below what its written documents presume. Bendick and Cantu (1978) examined how well the literacy levels of welfare clients matched the readability levels of 81 documents used in various income assistance programs. They found that most of the documents they examined were much more difficult than what their readers could handle. Even highly literate citizens have been known to have difficulty understanding Internal Revenue Service regulations. Research on text structure and comprehension has contributed to a movement toward plain English that has been joined by government agencies and private industry. The plain English movement attempts to develop and use principles of comprehension to revise the instructions and forms used in government and public service industries. One of the interesting aspects of the movement is that it represents basic research applied to real-world problems.

Readability

In the past, readability formulae provided an estimate of how easily a typical person would be able to read a particular text. The most common formulae, such as the Flesch Reading Ease Scale (1948) or the Dale-Chall formula (1948), use two factors to predict readability. One factor is the average number of words per sentence. The second factor is some measure of how common the individual words are--for example, their length (short words are more common) or their presence on a list of common words. The formulae try to predict what grade level of reading ability a student must have in order to understand a given text. While readability formulae like these may have some usefulness in identifying difficult prose, they are not very helpful in revising complex instructions. Most readability formulae were originally developed to describe how readable textbooks were for school children, so they may not apply to other material, such as instructions and government forms, or other groups of readers, such as adults. Finally, the two variables of sentence length and vocabulary are simply too limited to describe the complex processes that determine comprehensibility (Bormuth, 1969:45). Readability formulae ignore structural aspects of prose beyond the sentence, such as the way sentences are put together into paragraphs. They ignore the background of the readers. They ignore the purpose of the document and the compatibility of the writing to the reader's goals. For instance, one problem with poorly written instructions is that the reader often cannot find the particular information that he or she is seeking in a document. On a practical level, readability formulae have undue influence on publishers simply because they are so easy to use. The fact that a text is incoherent may be ignored because it passes some limited formula for its level of readability.

Revising a text just by improving its readability rating generally does make it substantially more comprehensible. Klare (1976) reviewed 36 experimental studies that tested the effect of readability variables on how well the readers comprehended and remembered the material, and only about half the studies showed an effect. Those that did produce an improvement in comprehension and memory required a very large change in the materials, a difference equivalent to 6.5 grade levels, such as changing a story from a 9th-grade to a 2nd-grade level. Similarly, a study by Kern (1979) found that changing vocabulary and sentence length in material given to Navy personnel generally did not alter its comprehensibility. Even if the documents are changed so that they meet the requirements of the formulae, there is no guarantee that they will be readable.

A recent goal of language research is to develop a scientific basis for writing comprehensible texts. The scientific issue is to determine how language structures interact with psychological processes. The general theory is that certain texts are less readable because they require the reader to frequently search long-term memory for the information necessary to interpret what is currently being processed (Kintsch and Vipond, 1979). This approach has been able to successfully predict the readability of some passages, when the criterion of readability was related to comprehension measures. Other research has shown how various syntactic devices can facilitate reading by directing the reader's attention to the major thematic foci of a passage (Carpenter and Just, 1977). A related approach is to determine what cues readers use to distinguish important and unimportant information (Kieras, in press). For example, the opening sentence of a paragraph is often interpreted as the theme, even when it is not. This approach attempts to construct detailed models of what makes a text comprehensible by examining how readers interpret it.

A closely allied area of research has made a direct attempt to examine how government and public documents can be improved. One example is a study that looked at how readers "translate" a typical document to themselves as they read it (Flower et al., 1980). The document was from the regulation governing the Small Business and Capital Ownership Development Program, which is administered by the Small Business Administration. The readers were three people in business who were likely to encounter such regulations. The readers often revised the wording for their own benefit as they read. One of the most striking findings is that the readers constructed scenarios, short stories, or examples to clarify a statement. The readers tried to express the regulations concretely, in terms of what a person should do.

Expert writers also try to write in terms of actions. In one case expert writers revised a regulation that has been praised as easy to read (the Health Education Assistance Loan regulation). The writers used more people-centered statements than did the Small Business Administration writers. For example, the statement "A student must be: a citizen, national or permanent resident in the United States" is an improvement over the old regulation, which often began, "Eligibility is . . . " or "Determinations will be made . . . " (Flower et al., 1980). Clearly written documents also focus on the information that the reader will need in order to act, rather than focusing on definitions and abstract information.

A document that has a number of examples and cases can become lengthy. In addition, readers sometimes mistakenly assume that the examples represent the entire range of a term's meaning. But for certain documents, they may be the best approach. For example, the Swedish income tax authorities write their instructions as a series of elaborate scenarios that cover the most typical cases. Because of the clear style, most Swedes fill out their own forms. People who do not conform to the given examples must read the more complex instructions (Flower et al., 1980). When a few examples will cover most of the relevant cases, instructions in the form of concrete examples seem to be worthwhile. In sum, it appears that one outcome of current research in reading will be a model of how to make written documents more comprehensible.

SUMMARY

We began this paper by arguing that reading comprehension is an increasingly important skill for American society. Fewer jobs are available for people who do not have general literacy skill, and the outlook is that such jobs will decrease. Within this context, there appear to be two issues concerning reading: It is important that citizens attain minimal reading skill and, beyond this, that they attain fluency in reading to the level of their oral comprehension. We have examined what basic research has learned thus far that may contribute to these dual goals.

Basic research in several disciplines has led to a new interest in reading. In a relatively short time, there has been considerable success in formalizing what is known about the structure of language, the nature of the perceptual process in reading, sources of individual differences, the interaction of multiple processes in the skilled reader, and some aspects of reading acquisition in children. At the same time, research has led to and been accompanied by concern for some of the central educational issues, such as the nature and sources of dyslexia, the reading problems of the deaf, and how to improve communication not only by changing the reader but also by changing the way material is written. A basic argument we have made is that reading is based on language skills and general knowledge. Consequently, reading problems sometimes reflect problems with language understanding more generally, not just with the written word. These distinctions, which are being refined and supported by basic research, may have large implications for how reading is viewed by society and remediated in the schools. It suggests that reading must be viewed in the context of spoken language and general problem solving.

REFERENCES

Ahlstrom, K. G.
1971 "On evaluating the effects of schooling." Proceedings of the International Congress on Education of the Deaf, Stockholm.

Anderson, T. H., B. B. Armbruster, and R. N. Kantor
1980 How Clearly Written Are Children's Textbooks? Or, of Bladderworts and Alfa. Reading Education Report No. 16. University of Illinois at Urbana-Champaign.

Baron, J.
1977 "Mechanisms for pronouncing printed words: use and acquisition." In D. LaBerge and S. J. Samuels, eds., Basic Processes in Reading: Perception and Comprehension. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Baron, J., and I. Thurston
1973 "An analysis of the word superiority effect." Cognitive Psychology 4:207-228.

Beck, I. L.
1981 "Reading problems and instructional practices." In T. G. Waller and G. E. MacKinnon, eds., Reading Research: Advances in Theory and Practice. Volume 2. New York: Academic Press.

Bellugi, U., E. S. Klima, and P. Siple
1974-1975 "Remembering in sign." Cognition 3:93-125.

Bendick, M., Jr., and M. G. Cantu
1978 The Literacy of Welfare Clients. Washington, D.C.: The Urban Institute.

Benton, A. L., and D. P. Pearl, eds.
1978 Dyslexia: An Appraisal of Current Knowledge. New York: Oxford University Press.

Birch, H. G.
1962 "Dyslexia and maturation of visual function." In J. Money, ed., Reading Disability: Progress and Research Needs in Dyslexia. Baltimore: Johns Hopkins University Press.

Bloom, B. S.
1976 Human Characteristics and School Learning. New York: McGraw-Hill.

Bormuth, J. R.
1969 Development of Readability Analyses. University of Chicago Final Report, Project No. 7-0052. Washington, D.C.: U.S. Office of Education.

Bradshaw, J. L.
1975 "Three interrelated problems in reading: a review." Memory and Cognition 3:123-134.

Bransford, J. D., and M. K. Johnson
1973 "Considerations of some problems of comprehension." In W. G. Chase, ed., Visual Information Processing. New York: Academic Press.

Brown, A. L.
1980 "Metacognitive development and reading." In R. J. Spiro, B. C. Bruce, and W. F. Brewer, eds. Theoretical Issues in Reading Comprehension: Perspectives from Cognitive Psychology, Linguistics, Artificial Intelligence, and Education. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Brown, A. L., and S. S. Smilely
1977 "Rating the importance of structural units of prose passages: a problem of metacognitive development." Child Development 48:1-8.

Bryant, N.
1968 "Some principles of remedial instruction for dyslexia." In G. Natchez, ed., Children with Reading Problems. New York: Basic Books.

Carpenter, P. A., and M. A. Just
1977 "Integrative processess in comprehension." In D. LaBerge and S. J. Samuels, eds., Basic

Processes in Reading: Perception and Comprehension. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Chall, J.
1967 Learning to Read: The Great Debate. New York: McGraw-Hill.

Charniak, E.
1972 "Toward a model of children's story comprehension." TR-266. MIT Artificial Intelligence Laboratory, Cambridge, Mass.

Chomsky, N.
1957 Syntactic Structures. The Hague: Mouton.

Conrad, R.
1977 "The reading ability of deaf school-leavers." British Journal of Educational Psychology 47:138-148.

Curtis, M. E.
1981 "Development of components of reading skill." Journal of Educational Psychology.

Dale, E., and J. S. Chall
1948 "A formula for predicting readability." Educational Research Bulletin 27:11-20, 37-54.

Denckla, M. B., and R. Rudel
1976a "Naming of object drawings by dyslexia and other learning disabled children." Brain and Language 3:1-15.

Denckla, M. B., and R. Rudel
1976b "Rapid 'automatized' naming (R.A.N.): dyslexia differentiated from other learning disabilities." Neuropsychologia 14:471-479.

DiFrancesca, S.
1972 Academic Achievement Test Results of a National Testing Programme for Hearing-Impaired Students. United States: Spring 1971, Report No. 9, Series D. Washington, D.C.: Gallaudet College, Office of Demographic Studies.

Fisher, D.
1978 Functional Literacy and the Schools. Washington, D.C.: National Institute of Education.

Flesch, R. F.
1948 "A new readability yardstick." Journal of Applied Psychology 32:221-233.

Flesch, R.
1955 Why Johnny Can't Read and What You Can Do About It. New York: Harper & Brothers.

Flower, L. S., J. R. Hayes, and H. Swarts
1980 Revising Functional Documents: The Scenario Principle. Document Design Center Report. Pittsburgh, Pa.: Carnegie-Mellon University.

Fry, M. A.
1967 "A transformational analysis of the oral language used by two reading groups at the second grade level." Unpublished doctoral dissertation. University of Iowa.

Gibson, E., and H. Levin
1975 The Psychology of Reading. Cambridge, Mass.: MIT Press.

Hall, W. S., and R. Freedle
1975 Culture and Language. New York: Halsted Press.

Hall, W. S., and L. F. Guthrie
1980 "On the dialect question and reading." In R. J. Spiro, B. C. Bruce, and W. F. Brewer, eds., Theoretical Issues in Reading Comprehension: Perspectives from Cognitive Psychology, Linguistics, Artificial Intelligence, and Education. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Hall, W. S., S. Reder, and M. Cole
1975 "Story recall in young black and white children: effects of racial group membership, race of experimenter, and dialect." Developmental Psychology 11:828-834.

Hardyck, C. D., and L. F. Petrinovich
1970 "Subvocal speech and comprehension level as a function of the difficulty level of reading material." Journal of Verbal Learning and Verbal Behavior 9:647-652.

Hatfield, N., F. Caccamise, and P. Siple
1978 "Deaf students' language competency: a bilingual perspective." American Annals of the Deaf 123:847-851.

Hunt, E., C. Lunneborg, and J. Lewis
1975 "What does it mean to be high verbal?" Cognitive Psychology 7:194-227.

Ingram, T. T. S., and J. F. Reid
1956 "Developmental aphasia observed in a department of child psychiatry." Archives of Disorders of Childhood 31:161.

Ingram, T. T. S., A. W. Mason, and I. Blackburn
1970 "A retrospective study of 82 children with reading disability." Developmental Medicine and Child Neurology 12:271-281.

Jackson, M. D., and J. L. McClelland
1979 "Processing determinants of reading speed." Journal of Experimental Psychology: General 108:151-181.

Jenkins, J., and A. Liberman
1972 "Background to the conference." In J. F. Kavanaugh and I. G. Mattingly, eds., Language by Ear and by Eye. Cambridge, Mass.: MIT Press.

Just, M. A., and P. A. Carpenter
1980 "A theory of reading: from eye fixations to comprehension." Psychological Review 87:329-354.

Kern, R. P.
1979 Usefulness of Readability Formulas for Achieving Army Readability Objectives: Research and State-of-the-Art Applied to the Army's Problem. Fort Benjamin Harrison, Ind.: Technical Advisory Service, U.S. Army Research Institute.

Kieras, D. E.
in press "The role of major referents and sentence topics in the construction of passage macrostructures." Discourse Processes.

Kintsch, W., and T. A. van Dijk
1978 "Toward a model of text comprehension and production." Psychological Review 85:363-394.

Kintsch, W., and D. Vipond
1979 "Reading comprehension and readability in educational practice and psychological theory." In L. G. Nilsson, ed., Perspectives on Memory Research. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Klare, G. R.
1976 "A second look at the validity of readability formulas." Journal of Reading Behavior 8: 129-152.

Klima, E. S., and U. Bellugi
1979 The Signs of Language. Cambridge, Mass.: Harvard University Press.

Labov, W.
1970 "The logic of non-standard English." In F. Williams, ed., Language and Poverty. Chicago: Markham.

Lerner, B.
1981 "The minimum competence testing movement: social, scientific, and legal implications." American Psychologist 36:1057-1066.

Liberman, I. Y., D. Shankweiler, C. Orlando, K. S. Harris, and F. B. Berti

1971 "Letter confusion and reversals of sequence in the beginning reader: implications for Orton's theory of developmental dyslexia." Cortex 7: 127-142.

Liberman, I. Y., D. Shankweiler, A. M. Liberman, C. Fowler, and F. W. Fischer

1977 "Phonetic segmentation and recoding in the beginning reader". In A. S. Reber & D. L. Scarborough, eds., Toward a Psychology of Reading: The Proceedings of the CUNY Conferences. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Louis Harris and Associates

1970 Survival Literacy: Conducted for the National Reading Council. New York: Louis Harris and Associates.

Lyle, J. G.

1970 "Certain antenatal, perinatal, and developmental variables and reading retardation in middle class boys." Child Development 41:481-491.

Mandler, J. M., and N. S. Johnson

1977 "Remembrance of things parsed: story structure and recall." Cognitive Psychology 9:111-151.

Markman, E. M.

1977 "Realizing that you don't understand." Child Development 48:986-992.

McClelland, J. L., and D. E. Rumelhart

1981 "An interactive activation model of context effects in letter perception: Part 1. An account of basic findings." Psychological Review 88:375-407.

Meyer, D. E., and R. W. Schvaneveldt

1971 "Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations." Journal of Experimental Psychology 90:227-234.

Miller, G. A.

1973 Linguistic Communication: Perspectives for Research. Newark, Del.: International Reading Association.

Moores, D. F.

1971 An Investigation of the Psycholinguistic Functioning of Deaf Adults. Research Report No. 18. Washington, D.C.: Bureau of Education for the Handicapped, Office of Education, U.S. Department of Health, Education, and Welfare.

Moores, D. F.
1972 "Neo-oralism and the education of the deaf in the Soviet Union." Exceptional Children 38: 377-384.

Murphy, R. T.
1973 Adult Functional Reading Study. PR 73-48. Princeton, N.J.: Educational Testing Service.

Murphy, R. T.
1975 Adult Functional Reading Study. PR 75-2. Princeton, N.J.: Educational Testing Service.

National Assessment of Educational Progress
1976 Reading: Summary. Report 02-R-00. Education Commission of the States, Denver, Colo.

Newell, A., J. C. Shaw, and H. A. Simon
1957 "Empirical explorations of the logic theory machine: a case study in heuristics." Proceedings of the Joint Computer Conference: 218-230.

Perfetti, C. A., and T. Hogaboam
1975 "The relationship between single word decoding and reading comprehension skill." Journal of Educational Psychology 67:461-469.

Perfetti, C. A., and A. M. Lesgold
1977 "Discourse comprehension and sources of individual differences. In M. A. Just and P. A. Carpenter, eds., Cognitive Processes in Comprehension. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Pichert, J. W., and R. C. Anderson
1977 "Taking different perspectives on a story." Journal of Educational Psychology 69:309-315.

Rayner, K.
1975 "The perceptual span and peripheral cues in reading." Cognitive Psychology 7:65-81.

Reddy, D. R.
1980 "Machine models of speech perception." In R. A. Cole, ed., Perception and Production of Fluent Speech. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Resnick, D. P., and L. B. Resnick
1977 "The nature of literacy: an historical exploration." Harvard Educational Review 47:370-385.

Resnick, L. B., and P. A. Weaver, eds.
1979 Theory and Practice of Early Reading. Volumes 1, 2, 3. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Rodenborn, L. V., and E. Washburn
1974 "Some implications of the new basal readers." Elementary English 51:885-888.

Rozin, P., and L. R. Gleitman
1977 "The structure and acquisition of reading II: the reading process and the acquisition of the alphabetic principle." In A. S. Reber and D. L. Scarborough, eds., Toward a Psychology of Reading: The Proceedings of the CUNY Conferences. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Rumelhart, D. E.
1975 "Notes on a schema for stories." Pp. 211-236 in D. G. Bobrow and A. M. Collins, eds., Representations and Understanding: Studies in Cognitive Science. New York: Academic Press.

Rutter, M.
1978 "Prevalence and types of dyslexia." In A. L. Benton and D. Pearl, eds., Dyslexia. New York: Oxford University Press.

Schank, R. C., and R. P. Abelson
1977 Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Schulte, C.
1967 "A study of the relationship between oral language and reading achievement in second graders." Unpublished Ph.D. dissertation. University of Iowa.

Spilich, G. J., G. T. Vesonder, H. L. Chiesi, and J. F. Voss
1979 "Text processing of domain-related information for individuals with high and low domain knowledge." Journal of Verbal Learning and Verbal Behavior 18:275-290.

Stein, N. L., and T. Trabasso
1981 "What's in a story: critical issues in comprehension and instruction." In R. Glaser, ed., Advances in the Psychology of Instruction. Volume 2. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Sternberg, S.
1969 "Memory scanning: mental processes revealed by reaction-time experiments." American Scientist 57:421-457.

Stewart, W. A.
1969 "On the use of Negro dialect in the teaching of reading." In J. C. Baratz and R. Shuy, eds., Teaching Black Children to Read. Washington, D.C.: Center for Applied Linguistics.

Sticht, T. G.
1972 "Learning by listening." In J. Carroll and R. Freedle, eds., Language Comprehension and the Acquisition of Knowledge. Washington, D.C.: V. H. Winston and Sons.

Sticht, T. G.
1975 "The acquisition of literacy by children and adults." In F. B. Murray and J. J. Pikulski, eds., The Acquisition of Reading. Baltimore: University Park Press.

Sticht, T. G.
1979 "Developing literacy and learning strategies in organizational settings." In H. F. O'Neil, ed., Learning Strategies: Issues and Procedures. New York: Academic Press.

Sticht, T. G., J. S. Caylor, R. P. Kern, and L. C. Fox
1972 "Project REALISTIC: determination of adult functional literacy levels." Reading Research Quarterly 7:424-465.

Stuckless, E. R., and J. W. Birch
1966 "The influence of early manual communication on the linguistic development of deaf children." American Annals of the Deaf 111:452-460, 499-504.

Tierney, R. J., J. Mosenthal, and R. N. Kantor
1980 Some Classroom Applications of Text Analysis: Toward Improving Text Selection and Use. Reading Education Report No. 17. University of Illinois at Urbana-Champaign.

Vellutino, F. R.
1977 Dyslexia: Theory and Research. Cambridge, Mass.: MIT Press.

Vellutino, F. R., H. Smith, J. A. Steger, and M. Kamin
1975 "Reading disability: age differences and the perceptual-deficit hypothesis." Child Development 46:487-493.

Vestberg, R.
1973 "Evaluation of reading achievements of deaf children." In E. Kampp, ed., Evaluation of Hearing Handicapped Children. Denmark: Ebeltoft.

Wheeler, D.
1970 "Processes in word recognition." Cognitive Psychology 1:59-85.
Winograd, T.
1972 "A program for understanding natural language." Cognitive Psychology 3:1-191.

Territory, Property, and Tenure

Robert McC. Netting

Principles of property holding are often regarded as a subfield of law or a dry branch of the rather arid study of legal history. Unless one is involved in a boundary dispute with a neighbor or in a political controversy over rights to graze or mine or drive off-road vehicles on federal wilderness land, the question of who owns what may seem rather abstract and academic. Even conflict over property can be resolved, we feel, once the facts of the case are known and the appropriate legal rules judiciously applied. But like so many cultural premises we take for granted, rights to resources are themselves problematic. Because the variety of understandings about holdings in land or other goods is tremendous when viewed cross-culturally, confining our view to our own and similar codified land law is both ethnocentric and unscientific. A purely descriptive or a legalistic approach to land tenure also prevents the development of theory on how this crucial aspect of human society functions and how it developed through historic time.

The social sciences are challenging our unquestioned, conventional views of ownership, both by analyzing non-Western concepts and behavior concerning rights to resources and by investigating troublesome cases, such as the collision of private property and eminent domain in our own society. Comparative law and jurisprudence have long pursued such concerns, but, except for the early seminal thinkers such as Montesquieu and Maine, the tendency was often to encapsulate land law, customs of inheritance, and contracts as areas artificially cut off from other kinds of social activity and guarded by scholar-specialists. The modern thrust of the social sciences has been frankly interdisciplinary and oriented toward common problems rather than the defense of intellectual turf.

Land tenure, I contend, does not make sense unless considered as part of a system involving the products of that land, the technology applied to gain subsistence, and the population sustaining itself from these resources. The grand tradition of economics dating from Malthus and Adam Smith was occupied in part with questions of how changes in production, tools, and competition for resources affect the market price of land, its consolidation or fragmentation, and the conditions of tenancy and rent. Land for Marx was one of the guises of capital, distinguished decisively from an earlier or more primitive stage, when rights to resources were held in common. Modern agricultural economists have been deeply involved in defining the range of tenure types under various systems of exploitation (e.g., plantations, ranches, truck gardens, diversified family farms, agribusinesses) and calculating the costs and benefits of each. With anthropologists, political scientists, and rural sociologists, they have begun to uncover the special characteristics of land use by what are called shifting cultivators or nomadic pastoralists in the less developed countries. Such studies inevitably involved the natural sciences as well. Geography brought the contributions of geology, soil science, meteorology, and botany to bear on questions of why one form of subsistence prevailed rather than another. Systems theorists and ecologists examined the distinctive input and output flows of energy. Both the use of resources and the rights socially assigned to them came to seem more complex and less arbitrary than many had suspected. Anthropologists, geographers, and economists together arrived at notions of peasant rationality and the strategies by which self-sufficient cultivators minimized risk rather than maximizing a particular kind of output (Lipton, 1968). Decisions to sharecrop rather than rent or buy a field were shown to be carefully calculated and frequently optimal (Johnson, 1971; Scott, 1976). The flexibility in actual practices of landholding in contrast to the strict legal definitions have been admirably demonstrated by social historians using voluminous early court records (Raftis, 1964). Students of contemporary economic development have now compiled extensive comparisons of land reform programs in many parts of the world (de Janvry, 1980; Land Tenure Center, 1974). Synthesizing the results of such widely proliferating basic research is not practical, but it is possible to point out an emerging consensus on land tenure issues of particular significance to anthropologists.

PROPERTY AND THE EVOLUTIONARY PARADIGM

> The idea of property was slowly formed in the human mind, remaining nascent and feeble through immense periods of time. Springing into life in savagery, it required all the experience of this period and of the subsequent period of barbarism to develop the germ, and to prepare the human brain for the acceptance of its controlling influence. Its dominance as a passion over all other passions marks the commencement of civilization. It not only led mankind to overcome the obstacles which delayed civilization, but to establish political society on the basis of territory and of property. A critical knowledge of the evolution of the idea of property would embody, in some respects, the most remarkable portion of the mental history of mankind (Morgan, 1963:5-6, orig. 1877).

It is no surprise that Lewis Henry Morgan, the Rochester lawyer, railroad investor, and New York state legislator--as well as the father of American anthropology--was interested in ideas of property. In him the enthusiasm of the Victorian amateur scientist led to both an exploration of the structure of the Iroquois Confederacy and to a defense of the Seneca from a threatened land grab (Leacock, 1963). For Morgan, the great watershed of human society distinguishes those smaller, presumably ancestral groups, which are organized on the basis of kinship, clans, and tribes, from kingdoms and nations, which are "founded upon territory and upon property." Such grand evolutionary dichotomies have a disconcerting way of evaporating or being qualified out of existence by the evidence of the careful, firsthand studies of unfamiliar societies that are the distinctive contribution of modern anthropology. Yet the ethnological investigations of religion, material culture, descent groups, and language that followed Morgan did not deal with property as a central concern. We were told about the ceremonies conducted when one crossed a tribal border (Van Gennep, 1960) or the destruction of canoes, coppers, and other valuable pieces of property in a Northwest Coast potlatch (Benedict, 1934), but only more recently has a theory of property or rights to resources begun to emerge.

Studies of hunting territories, the ownership of tools and dwellings, and land tenure were neglected, perhaps because Morgan's principle seemed self-evident. Technologically primitive peoples, gathering naturally occurring foods from their surroundings and sharing the products of the chase, would have no need to establish personal rights to anything, whereas farmers and herders could protect their claim to scarce land or domestic animals only by an assertion of private property. Morgan's followers insisted that individual property in simple societies was purely personal, while land, the basic source of subsistence, was always collectively held (Leacock, 1963:xvi). Opponents of the evolutionary view found private ownership in every society as well as an ethic of sharing and well-defined communal rights belonging to a variety of groups, such as households, lineages, clubs, villages, and tribes (Lowie, 1920). Perhaps because Morgan's work was used by Engels in The Origin of the Family, Private Property and the State, critics decried the "dogma of a universal primitive communism" (Lowie, 1920:235). Empirical questions about the nature and function of property have been translated into the fighting words of political parties and contending philosophies of government. It is ironic that Morgan's (1963) insights on the role of property in human affairs, along with his oracular vision that "human intelligence will rise to the mastery over property, and define the relations of the state to the property it protects, as well as the obligations and the limits of the rights of its owner" (p. 561), have served less to clarify the issues involved than to remove them from the realm of objective scientific analysis.

Although the confidence of the 19th-century evolutionists that they could discern in human history a single progressive development from undifferentiated, amorphous, communal rights to specific, legally sanctioned, heritable, private property was misguided, the questions of who uses what and who owns what are still central ones. By recognizing the fallacy of some simple and necessarily ethnocentric definition of the nature of property and ownership, anthropologists have opened the matter for basic research. As long as property rights were considered a branch of law, with conflicting claims to be settled only by reference to courts and statutes, the cross-cultural experience of different societies having contrasting systems of subsistence was irrelevant. The single most important advance in the study of property has been the recognition that rights to resources could not be understood apart from the human ecosystem in which they existed.

Hunter-gatherer territoriality is meaningful in terms of climate, topography, precipitation, the movements of game animals, the distribution of edible plants, and the tools and knowledge by which these resources are used to support human life. Relative scarcity of agricultural land in relation to population, the organization of farming labor, the efficiency of implements, and the productivity of crops must all be considered in order to fathom the rules and practices of land tenure. Cultural premises, perhaps residing in venerable tradition or new legislation or altered by conquest, must also be consulted, because what people do with property is dependent in part on what they think about it, the cultural concepts of what is appropriate and what is contrary to custom. It is axiomatic, then, that the ecological study of property is interdisciplinary. If I favor the results from anthropology, it is merely because I have more familiarity with them; the approach owes much to biology, agricultural economics, geography, demography, and social history as well.

Anthropology, though striving for functional interpretations or the modeling of behavioral systems, is wary of wide-ranging generalizations or abstract deductions. The ethnography of a particular society observed at a specific place and time is the raw material used by anthropologists, and our hypotheses tend to be cautious and middle-level. Since they often grow, as it were, inductively from field work, I rely heavily in this paper on case studies, attempting to test these approximations through controlled comparisons of similar societies and analyses of change in property or land tenure practices through time in a single society. I am concerned largely with questions that many social scientists are asking but for which the answers remain matters of dispute. The ethnographic studies cited cover areas of consensus and disagreement on hunter-gatherer territoriality, fishing rights to bodies of water, and landholding among shifting and intensive agriculturalists. In most of these cases, the same processes can be charted in simple, self-sufficient societies as in sections of modern society using parallel techniques of production. Issues of communal versus individual rights recur as we go from the desert water hole and the bush field to the offshore fishing limit in international waters and the claims of the Sagebrush Rebellion. Three themes dominate the discussion of illustrative cases: (1) The use of the land or the sea by and large determines the rights that people exercise over land or sea resources. Ownership or tenure cannot be divorced from the ecological situation of environment, technology, population, and social organization. (2) A change in the ratio of population to available resources may change the competition for these resources and the equality of access to them. While rules of land tenure remain the same, the distribution of important resources within the population may undergo significant and predictable change, with the proportion of landless people increasing as the population grows. (3) A lack of understanding of the conceptions and operations of property systems in other societies is a frequent cause of conflict, injustice, and exploitation. More adequate analysis of historic and existing modes of tenure would have implications for policy regarding land reform and controversies involving claims to individual versus communal rights to property.

HUNTER-GATHERER TERRITORIALITY

From antiquity, philosophers have used the idealized or imagined societies of nomadic hunter-gatherers as examples of some primeval state of human nature. Where tools were simple and easily fabricated and where daily food was gained directly from the environment, it was plausibly surmised, individual possessions would be few, sharing would be frequent, and there would consequently be little differentiation in wealth or property among people. For some this meant that primitives lacked any concept of ownership and moved at will across the landscape (Herskovits, 1952:331). Other authors speculated that cooperation within groups would be accompanied by hostility between groups. Each band or horde or tribe would "own" hunting grounds or specific resources, such as berry patches, fruit trees, and water supplies, that would be defended with force against trespass by neighbors. Since the local group was often structured by kinship ties, it was often the clan or the patrilineal band that was believed to exercise exclusive rights to a territory. During the great age of European overseas exploration and expansion, it was further assumed by those interested in acquiring farming or ranch lands or removing troublesome indigenes that native rights to bounded tracts could be directly acquired by sale or treaty from the legal representatives of the group. Subduing the wilderness and planting crops was also felt to be a morally commendable, religiously sanctified land use that justified the appropriation of untilled "wasteland."

Given the theoretical emphasis placed by the evolutionists on living examples of purportedly earlier or elementary forms of human social life, it is interesting that basic research on hunter-gatherers is relatively recent. Territorial rights could not be elucidated until anthropologists stopped asking individuals inappropriate questions, such as what land their group owned or where the tribal boundaries were located. Only when an ecological approach was adopted, which included detailed descriptions of real resource use, movements, work group cooperation, residence history, local kin ties, and intergroup conflict, could the material correlates of ideas of property be assessed.

Shoshonean Territories

Julian Steward's work (1938, 1955) on the Paiute and Shoshonean Indians of the Great Basin, in what is now Nevada, Utah, and portions of California, investigated territoriality as Steward laid the foundation for what he called cultural ecology. Unpredictable rainfall and little standing water in the high valleys meant that vegetation was sparse and scattered and game animals were mobile and seldom concentrated in herds. The inhabitants used seed beaters and baskets to gather grass seeds; bows, clubs, and nets to kill jackrabbits, antelope, and deer; and subsisted on pine nuts from the pinyon groves in winter. With such a spotty and variable occurrence of resources, people lived in small camps near water and moved frequently. They roamed large stretches of countryside, assembling for game drives but often dispersing into family groups because their subsistence activities did not benefit from large-scale cooperation. Winter settlements depended on the size and location of the pinyon harvest. Under such circumstances, population density, the size, composition, and movements of productive groups, and rights to the means of production are what Steward called "culture core" features closely related to subsistence activities and economic arrangements (Netting, 1965a).

As is usually the case with hunter-gatherers, the Great Basin people had personal tools, and the labor one expended on gathering wild plant foods gave one rights to consume or distribute them. People moved from camp to camp and often visited neighboring groups in which they had kin. Rigid or exclusive rights to territories did not exist (Steward, 1977:375):

> Social groups in the Basin-Plateau area obviously had to confine themselves to familiar territory, for the knowledge of the location of resources, including water, precluded indefinite and random wandering in strange terrain. That a group exploited about the same territory each year, however, did not imply exclusive claims to or defense of its resources. An informant's statement that "this was our territory; we owned it" is almost invariably followed by the further statement that anyone was free to use the resources. In fact, while the small family clusters of Western Shoshoni traveled during the summer, they exchanged information with other clusters concerning the whereabouts of seeds and game and especially about the prospects for the pine nut harvest in different mountains. The Kaibab Southern Paiute claimed . . . that each family cluster owned the watering place to which it customarily returned each winter, but this meant no more than that a cluster of families made the watering place its preferred headquarters or principal encampment.

Bands and Environmental Spacing

Though Steward effectively characterizes the Great Basin territories as familiar home ranges with vague, porous boundaries and without exclusive use rights, he believed that many hunter-gatherers were organized in patrilineal bands with local exogamy and land ownership (Steward, 1955:122). Competition for game would lead to conflict, so patrilineally related men would remain in the territory in which they were reared and band together to protect their game resources (Steward, 1955:135). This formulation has become increasingly suspect as the basic research of the 1960s and 1970s has demonstrated the flexibility of band membership and the ease with which hunter-gatherers move from one group territory to another (Lee and DeVore, 1968). Lee (1972) pointed out that among !Kung San of the Kalahari in Northern Botswana, a bilateral, nonterritorial grouping is a mechanism for responding to local imbalances in food resources and gives reciprocal access to resources that allows a much higher population density than could be supported if every territory had to contain a permanent water source. He argues that the patrilocal-territorial model of hunter-gatherer bands is empirically rare and unlikely, if not impossible, on theoretical grounds. But another student of the Bushmen (Silberbauer, 1972) contends that each band has a "resource nexus" over which its members have exclusive rights of exploitation and, furthermore, that certain individual "owners" are administrators of these rights on behalf of the band.

Some of these definitional ambiguities are clarified by looking at the more precise conceptions of territoriality developed by animal ecologists. The social behavior involved in staking out a specific volume of space for feeding or breeding and defending it against other members of the same species results in the adaptation of the organism to the available food supply and the maintenance of the population size at a reasonable maximum (Clapham, 1973:74-75). Territoriality may result from aggression or from attachment to a site or area. Australian aborigines, who were long thought to have classic patrilineal band organization, do not actively defend their boundaries (e.g., a 400-square-mile territory would have a 70-mile perimeter to defend), but a named local descent group with ritual sites and mythic, totemic ties to certain natural features is regularly associated with a particular area (Peterson, 1975). Territories are not rigid, and camps at any time may contain members from several clans and localities. The exigencies of the food quest and demographic variation among groups make such flexibility necessary. But everyone knows his or her native clan territory, and older clan members prefer to spend more of their time there. The ideology of kin group possession of an area is recognized when visitors ritually ask for formal acceptance into the group before they make use of local resources. Greeting ceremonies are functionally analogous to territorial defense, and a failure of visitors to announce their presence is regarded as an act of aggression (Peterson, 1975). Thus the continuing identification of a kin group with a territory and the optimal spacing of local populations in the environment are not inconsistent with the movement of individuals to make the best use of their resources. In fact, by utilizing a wide variety of ties through descent and marriage, people can claim admission to bands at some considerable distance from their own area of origin.

Though hunter-gatherers certainly show a seasonal variation in group size and fluidity of group composition that makes untenable the old law-and-order view of patrilineal territorial organization, there is a growing concern among ethnographers about how people in these nomadic societies think about rights to land (Peterson, 1979). The !Kung San have well-developed concepts of owner and territory: Rights come through either parent and are strongly held where one is resident; rights are residual or weak in the territory identified with the other parent (Marshall, 1976; Wiessner, 1977). Within a territory or n!ore, the water hole and area immediately around it is owned and inherited within a group of kin, but a surrounding broad belt of land is shared with adjacent groups (Lee, 1979:334). Publicly acknowledged rights to mongongo nut groves and wild bean patches are stated. Immigrants who join and stay with the group are gradually absorbed into the core identified as owners. Lee (1979:335) believes that the !Kung consciously strive to maintain a boundaryless universe (pp. 337-338):

> Although ownership is collective, not individual, and the boundaries are not well marked, the concept of ownership is there nonetheless. . . . Members of the core group and their visitors may exploit the resources of the n!ore without restriction. Neighboring groups may use the resources as well, but they must keep the owners informed of their movements and who they are camping with. More distant groups must ask permission more explicitly and should be modest in their behavior, in terms of length of stay and the number of people brought in. . . . Disputes between groups over food are not unknown among the !Kung, but they are rare, far less common and less serious than are, for example, fights over sex, adultery, and betrothal.

Desert Australians have a more elaborate religious ideology of land ownership, perhaps because of a more stable association between groups and permanent water sources, while the few !Kung water holes must be shared by a number of bands (Peterson, 1979). Similar factors may influence the greater territoriality of the !Ko San, who live in the south cental Kalahari, where there is no staple plant food and where melons and other plants are the sole source of water for part of the year. Plant resources are owned by the band, and permission must be requested to cross a well-defined boundary zone for hunting in another band's territory (Barnard, 1979). !Ko bands and band clusters are highly nucleated, endogamous, and dislike strangers, showing little of the !Kung flexibility in a less arid environment.

These examples may suggest that the scarcity, localization, and predictability of a resource along with the competition among groups for its use influence the degree to which property rights are asserted. The obvious adaptiveness of a home range of familiar territory for hunter-gatherers seems never to result in rigidly defended boundaries, but the crucial nature of a limiting resource such as water may encourage an ideological focus, identifying a defined group more closely with its territory. When the relationship of resources to population is altered, territorial claims should correspondingly tighten or relax. There has been a great deal of controversy about whether the hunting territories of Algonkian Indian groups in Canada were aboriginal (Hallowell, 1949; Leacock, 1954; Speck and Eiseley, 1939). In fact, family hunting territories were defined and sanctions against trespass were enforced where population density was highest and the competition for furs to trade at the Hudson Bay posts was most severe. When the Ojibwa population began to decline and big game such as moose returned to parts of Ontario around 1900, a distinct lessening of territoriality took place (Bishop, 1970). The assertion of ownership and the degree of exclusivity of use does not appear to be a constant or a purely customary value in a society but rather a set of beliefs and practices closely adjusted to other, potentially variable factors in a specific ecosystem.

FISHING RIGHTS TO SEA AND RIVER

Although notions of group territorial rights to hunting grounds and trap lines may be readily intelligible to people from property-based societies, we often imagine the oceans and streams to be held in common, with everyone enjoying unimpeded access to fish, crustaceans, and mollusks. The freedom of the seas implies not only rights of ship travel but also the global pursuit of tuna or whales. It is only in recent years that the codfish war off Iceland and the national claims to 200 miles of territorial waters, along with major declines in world commercial catches, have publicized territorial competition for marine resources. Several examples of traditional and modern fisherfolk demonstrate that especially productive fishing locations with high and seasonally dependable yields are often owned and forcibly defended by groups or individuals. These practices ensure a predictable subsistence for those who control the property as well as regulate overfishing and the rapid population declines to which marine resources are particularly subject.

Northwest Coast Fishing Sites

Among the Northwest Coast Indians who depended on the annual salmon runs for a major part of their subsistence, locations on rivers or estuaries, where the migrating fish could be effectively trapped in weirs, caught in reef nets, or speared, were owned by villages or localized kin groups. Claims were also made to shellfish beds, berry patches, and settlement sites. A Kwakiutl informant, George Hunt, spoke of kin groups owning rivers and fighting any trespassers who attempted to build a trap for salmon or olachen (Boas, 1921:1348). Ruthless warfare led to the appropriation of land, material wealth, and slaves as well as to the capture of crests and other ceremonial prerogatives that in effect gave title to these possessions. Sixteen named southern Kwakiutl political groups, each consisting of a set of households sharing a winter village, exploited the resources of a defined territory with its own salmon streams and acted as a unit having its own chief in potlatch and military affairs (Donald and Mitchell, 1975). The population of the local group correlated with the average productivity of its fishery, and larger groups had higher ranks in the potlatch hierarchy.

A variety of items of movable property, such as canoes, storage boxes, fishing equipment, blankets, and masks, were owned by individual families, while the head chiefs of extended families or lineages were the custodians or nominal owners of fishing spots and other types of fixed assets (P. Drucker, 1939). Potlatches, involving the giving away of food and wealth in return for recognition of the giver's social status (J. W. Adams, 1973), often included a public announcement of the tract from which the food had come and the lines of descent in which these property rights had been inherited (Drucker and Heizer, 1967). In this way, a statement of claim to valuable resources could be publicly witnessed by near neighbors who might be most likely to contest it. In the absence of written deeds or a centralized government to enforce property rights, the local group was officially restating its ownership of real property. Eating food with or accepting presents from a host group, who recounted their ancestral claims to productive property and ceremonially passed it on to a rightful heir, was equivalent to formally agreeing to respect these claims (Netting, 1977:36). When a corporate local group was declining in size or dying out, outsiders could become formal members by accepting the obligation of the potlatch for the group. Thus the potlatch also functioned to redistribute people among scarce and valuable localized resources while maintaining the system of legitimate property rights (J. W. Adams, 1981). Concepts of property on the Northwest Coast evidently allowed the inhabitants to move smoothly and successfully into a market economy, with a premium on acquisition and consumption, when European trade and colonization arrived (Codere, 1950).

Palauan Fish Traps and the Decline of Ownership

A change in the productivity and relative value of a marine location can alter the system of tenure applied to it. Reefs off the islands of Palau in Micronesia belong to the island communities and are loosely administered by their chiefs (McCutcheon, 1981). Certain deep holes in the lagoon coral are ideal spots for the setting of traps, and they were formerly claimed as property by the kebliil or nonunilineal clans. The clan chief issued use rights to a member of the clan, and the chief could also claim the best of the catch. As real property, these sites could be transferred like money or plots of land in traditional exchanges (McCutcheon, 1981:118). The introduction since 1950 of more productive nighttime fishing with spears, goggles, and underwater flashlights has made the fixed trapping spots outmoded. Fishermen can now follow the fish freely about the lagoon. Ownership of the trap sites and the former custom of asking permission to use them are no longer important. Furthermore, the Japanese government of Palau formally appropriated all parts of the sea below the high water mark, and successive governments have come to be regarded as the nominal owner of the reefs and the lagoon. Disputes over the use of marine resources can no longer be settled by local chiefs and elders. In this case, both changing uses and the imposition of rules by external authorities have operated to change Palauan rights to the sea.

Maine Lobstering: Property and Conservation

The marine environment has generally been regarded as common property, and it has been observed that this contributes to persistent overexploitation in many of the world's fisheries. "In the absence of ownership, fishermen have no incentives to curtail fishing activities in response to declines in catches or increases in costs, because no property right guarantees that fish not taken today will be available in large quantity or at greater weight in the future. What one fisherman does not catch today simply goes to the other fishermen" (Acheson, 1975: 183). Lobster fishing along the Maine coast is legally open to all who acquire licenses, but it is in fact divided informally into territories by groups of lobstermen who resist interlopers. Boundaries between territories are more precisely maintained in summer, when more people are fishing shallow areas near the shore, than in winter, when there is less competition for lobster found over a wider area of deep water. Violation of territorial boundaries may lead to warnings, followed by the cutting off of marker buoys and the deliberate removal of the offender's traps. Occasionally such incidents escalate into full-scale "lobster wars," in which groups of men destroy each other's traps and even boats (Acheson, 1975). Ownership of land on an island is held to mean ownership of fishing rights in nearby waters, even though ocean areas are legally part of the public domain. In areas in which boundaries are most clearly defined, island landowners may rent out fishing rights that are also inherited patrilineally.

The fact that some Maine lobstering areas are quite exclusive, or "perimeter-defended," while others have considerable overlapping zones of mixed fishing by men from several harbors allows a comparison of the production and conservation effects of two kinds of territoriality (Acheson, 1975:189-204). Where boundaries are more impermeable and better defended, there are fewer boats per square mile of productive waters, allowing a higher proportion of lobsters of the minimum legal size to remain uncaught and grow to larger sizes. Conservation is promoted in perimeter-defended territories by voluntary agreements among owners to limit the number of traps set and to request and abide by closed-season regulations. Survey data has shown that a higher percentage of larger lobsters and female lobsters reaching reproductive size come from perimeter-defended areas. Restrictions on the numbers of traps and fishing seasons also reduce lobster mortality. Biological benefits in perimeter-defended areas are accompanied by economic benefits to the lobstermen who catch more and bigger lobsters with less effort, resulting in substantially higher gross incomes than those of equally well-equipped and experienced fishermen in nucleated areas (Acheson, 1975:202-203).

Property that belongs to everyone, as Hardin (1968) has pointed out in his well-known analysis of the "tragedy of the commons," is often subject to overexploitation, as each user seeks maximum individual economic benefit by harvesting the largest possible share of the resource. Informal group ownership of the traditional perimeter-defended lobster territories controls access to the resource and prevents a competitive increase in fishing effort by enforcing local conservation measures (Acheson, 1975:205-206). A successful cooperative, such as a New Jersey group with a near monopoly of summer porgy fishing, can control production by imposing boat limits on themselves when market prices are low (McCay, 1981a). State and national governments may attempt to prevent abuse of marine resources by licensing fishermen, regulating the type of equipment and number of traps that may be used, and establishing closed seasons. Legalized fishing territories have not been established, in part because of political and economic barriers but also because rights of access to the "common" resources of the sea were conventionally thought of as a free good. In fact, the indigenous resource management systems of poor local fishermen may be overridden by the laws of wealthy modern nations that impose common property rights to marine resources and thus favor the highly capitalized, large-scale corporate fishing interests (McCay, 1981b).

Our cases from the Northwest Coast, Palau, and Maine suggest that where marine resources are highly productive and localized, there will be a tendency to establish group (not individual) exclusive property rights with defended boundaries, restrictions on group membership, and the voluntary regulation of use. These social constraints appear to adjust the populations of fishermen to the abundance of their prey and to promote the sustained yield of a fragile natural resource. It is by no means certain that higher levels of government regulating the management of local environments by statute and police power can preserve the flexibility and voluntary compliance that characterize the operations of small groups of independent producers (McCay, 1981a:372). As nation states of unequal size, populations, and technoeconomic development vie for the diminishing harvests of the sea, their vastly difficult international agreements may benefit from the lessons in fishery territoriality learned by societies whose life depends on effective ecological adaptation.

LAND TENURE IN SHIFTING CULTIVATION AND INTENSIVE AGRICULTURE

In the various kinds of foraging systems I have been examining, the problem has been one of how human populations establish rights in the area from which their food--in the form of wild plants, game animals, or fish--comes. I have dealt with the rights of small groups of resident producers as they hunted or gathered in areas with varying degrees of competition from other similar groups. The processes of possession I have outlined are rendered more explicit, elaborate, and pervasive as groups take up the rights of food producers to the crops, domestic animals, soil, water, and shelters that are integral to their distinctive life ways. In looking at farmers in the sections below, the old but ever-present dichotomy between collective, communal control and narrowly defined private ownership is refined and anatomized to more closely approximate the complexities of the real world.

Land tenure cannot be viewed as changing along some simple unilinear continuum from primitive to modern, but rather as an entity made up of a number of strands that are not always parallel. Rights can vary along three axes: (1) according to the length and specificity of the time over which they are exercised, from short-term, vague, or intermittent through specific and long-term to permanent; (2) according to the objects to which they are applied, whether few and limited, such as fruit trees, houses, or branded cattle, to inclusive and general, such as a land surface with all its products, surface and underground water, mineral resources, and air rights; and (3) according to the way in which they include and exclude other people as members of groups with different sorts of authority. Property involves social rights and duties with respect to objects of value and the specific sanctions that reinforce customary behavior (Hallowell, 1943).

A corollary of the rights of individuals and corporate groups to land would be the right to transmit property by inheritance, gift, loan, sale, conquest, or legal appropriation.

Is Property an Ecological Factor or a Mental Construct?

Previous discussions have taken an ecological approach in emphasizing the intimate systematic connections between territoriality and land use, resources, technology, and population. Such an essentially materialistic and empirical strategy runs counter to the priority of ideas and intellectual developments among certain legal and economic theorists. They believe that the evolution of exclusive property rights over the resource base was a prerequisite for the development of cultivation and domestication in the Neolithic Revolution (North and Thomas, 1970:241, cited in Runge and Bromley, 1979).

> When common property rights over resources exist, there is little incentive for the acquisition of superior technology and learning. In contrast, exclusive property rights which reward the owners provide a direct incentive to improve efficiency and productivity, or, in more fundamental terms, to acquire more knowledge and new techniques. It is this change in incentive that explains the rapid progress made by mankind in the last 10,000 years in contrast to his slow development during the era as a primitive hunter/gatherer.

Joint or communal property arrangments are thought to be inefficient, entailing transaction costs among users and inevitable resource depletion (Demsetz, 1967). A schematic model of the replacement of communal by individual private property rights neglects both different uses of land and the probability that a single society at one point in time may have several contrasting arrangements for land tenure adapted to quite different systems of land exploitation.

The new paradigm now developing in studies of land tenure rejects simple deterministic models of changes in land tenure as causes of changes in subsistence means or as direct results of legal codes or distinctive political philosophies. This development parallels the gradual disillusionment with models of evolutionary change in agriculture that emphasize technological improvement and the capture of energy as the prime and often sole motivating factors (L. A. White, 1959). It also appears that environmental determinism cannot account for a static agricultural adaptation by reference merely to soils, climate, and topography (Meggers, 1954). The more comprehensive framework emerging from the basic research of anthropologists, geographers, and economists links local increases in population density to agricultural intensification, in which higher labor input raises the total production of the land, with corresponding greater definition and more socially restricted holding of rights to land.

Anthropologists played a key role in explaining the workings of traditional systems of shifting or swidden cultivation, in which forest is cut and burned, crops are planted and harvested for one or a few years without benefit of the plow or animal traction, and the clearing is then allowed to revert to forest vegetation until its agricultural potential has been naturally regenerated (Conklin, 1957, 1961; Freeman, 1955; Geertz, 1963; Izikowitz, 1951). Under conditions of ample land and low population pressure, such slash-and-burn systems were shown to produce dependable crops with considerable efficiency of labor without permanent environmental degradation.

More intensive agriculture is characterized by keeping the field in crops for longer periods of time than it is fallow, through such techniques as crop rotation, manuring, irrigation, terracing, transplanting, more careful soil preparation, weeding, end fencing. Danish economist Ester Boserup (1965) pointed out that farming societies make the transition from shifting to intensive cultivation only when the local density of population forces them to use the land more continuously and effectively to maintain their desired level of food intake. Population growth, rather than being only a response to increased food production, as in the Malthusian view, could be thought of as a cause for agricultural change. People did not voluntarily adopt intensive methods because they were better or more efficient. Indeed, the labor required was increasingly thorough, complex, and arduous, and individuals, operating under a principle of least effort, were reluctant to do more work until the scarcity of land and other resources compelled them to. Although the Boserup view has been subjected to considerable criticism, emphasizing the variety of ways that a population can cope with or suffer from growth (Bronson, 1972; Cowgill, 1975;

Grigg, 1979; Nell, 1972), the general outlines of the model have received substantial empirical support (Clarke, 1966; Dumond, 1961; Gleave and White, 1969; Hanks, 1972; Netting, 1968, 1974; Spooner, 1972; Turner et al., 1977).

Shifting cultivation economizes on labor in a context of abundant land and easy access to resources. A swidden, once it has been burned and cropped, has rapidly declining yields due to weed competition, insect and animal pests, and the exhaustion of minerals contained in the ash (Moran, 1979). Shifting cultivators typically exercise rights in usufruct over the crops that they grow, but the fallow areas of regenerating bush have little immediate value. A clan or village may retain territorial claims to such tracts, and they are defended against incursions by neighboring groups, but only when fertility is restored are the plots reallocated on the basis of need to group members (Netting, 1969, 1977:75). Individuals vary in their requirements for fields according to the numbers of workers and dependents in their households and the labor on which they can call (Netting, 1965b). As long as there is more than enough land to go around, it makes sense to adapt use rights to demographic fluctuations while maintaining a community pool of fallow land for future exploitation.

Intensive Agriculture and Individual Rights

As land becomes scarcer and competition for its products increases, it is obvious that groups with more limited membership and eventually individuals should claim more continuing rights to its use. "The attachment of individual families to particular plots becomes more and more important with the gradual shortening of the period of fallow and the reduction of the part of the territory . . . not used in rotation" (Boserup, 1965:81). When the tribe or the community can no longer guarantee to a family land sufficient in quantity and quality for its needs, there is a tendency for the family to hold on to what it has. Intensification further increases the value of the land by the expenditure of labor for permanent improvements--fruit trees that continue to bear for years, manured gardens, terraces that prevent soil erosion, and wells or channels that allow dry season irrigation.

The correlation among factors of population density, agricultural intensification, and land tenure is visible both spatially and historically among the Ibo of eastern

Nigeria. The traditional system of tropical shifting cultivation was based on the growing of yams, cassava, and coco yams in areas of farmland that radiated from a central village like the sectors of a circle. A sector was cleared annually by the men of the village working together, and garden plots were then assigned to each extended family household. In those areas with severe population pressure of 400 to more than 1,000 people per square mile, the system has broken down and been replaced by the permanent cultivation of vegetables and tree crops using frequent fertilization and labor-intensive techniques of horticulture (Udo, 1965). Nuclear or small polygynous families have set up permanent residence in the midst of these farms, asserting both immediate possession and continuing ownership. Where farmland is in such short supply, the demand for it encourages arrangements for leasing it or mortgaging (pledging) it to someone in return for a sum of money (Netting, 1969). The area in which intensification and more individualized land tenure is evident does not differ markedly in crops or climate from other areas of Iboland in which shifting cultivation continues, but it is distinguished by its heavy local concentrations of population (Udo, 1965).

Though land tenure is easily conceived in terms of dichotomous communal and individual types, the reality of change reflects steps in a process. The Kofyar of the Jos Plateau in northern Nigeria have adjusted to population density and long-term limitations on available land by the intensive cultivation of terraced, ridged, and manured fields averaging 1.5 acres around every homestead. Boundaries are clearly marked and families who use a homestead and fields belonging to others must make an initial payment plus annual donations of palm oil, beer, chickens, and sometimes cash to the owner. Sons inherit land, houses, and livestock from their fathers, and the property of a man who dies without heirs reverts to the head of his lineage (Netting, 1968:170). The patrilineage does not control land as a group, but it can step in to prevent a member from selling or by other means permanently alienating a piece of land (p. 163). Outright sale of land, as found in the most crowded Ibo districts, appears to represent the last stage in the movement toward individual property rights. Among the Kofyar, wealth could be exchanged indirectly for land when a nonrelative donated the sacrificial cow or horse for the funeral commemoration feast of a man who died without heirs, thereby claiming rights to the dead man's homestead (pp. 170-171).

A controlled cross-cultural comparison of 15 ethnic groups in the New Guinea highlands supports the association of higher population density with greater agricultural intensity and more individual land tenure (Brown and Podolefsky, 1976). Intensive cultivation of sweet potatoes in this area makes use of fencing, drainage ditches, erosion control, and fertilizing. The closest correlation is between land tenure and the length of the fallow period: "All cases of . . . individual tenure are in societies where agriculture is permanent or fallow is less than six years; group tenure . . . is found with longer fallow periods. While group territory is recognized nearly everywhere, individual plots are held and inherited mainly where the fallow period is short and trees or shrubs are planted by the owner" (Brown and Podolefsky, 1976:221).

Conflict Over Land

More specific and definitive rights to land use, transfer, inheritance, and alienation seem to arise from the disputes that increase in number and severity as the supply of land becomes restricted. Even shifting cultivators may find that their existing territory is no longer adequate to support them. Rather than intensifying their agriculture, they may attempt to expand by conquering and expelling neighboring groups. Maori warfare has been attributed to this competition for forest fallow (Vayda, 1961). The segmentary lineages of the Tiv in the Nigerian savanna zone have endemic border arguments in which individuals extend their farms, then call out their kin groups for acrimonious debates on where the boundary should be. Since the outward push is characteristically directed against the most distantly related lineage or against a foreign ethnic group, a direction of movement is established, with each lineage losing land in the rear and gaining ground in front (P. Bohannan, 1954; Sahlins, 1961).

If warfare is not effective or migration is impeded, formerly temporary rights in land may be solidified and given legal standing by some regularized form of dispute settlement. With a growing population and the investments required for intensification, litigation over land should go up. Traditional political and religious authorities who are able to mediate such conflict among communities or kin groups and to resolve them without bloodshed perform a valuable social service and enhance their own prestige (Netting, 1972; Ottenberg, 1958). Although the process of tenure change may be almost unnoticed as the fallow period shortens and the period of usufruct becomes permanent (Netting, 1968:161), the inevitable conflicts focus public recognition on the precedents established by a chief or court as rights are adjudicated and the emerging law is enforced.

Kin Group Solidarity and Rights to an Estate

The social solidarity and cohesion of a descent group such as a unilineal lineage or a kindred is often regarded by anthropologists as axiomatic. Kinship is sui generis and cannot be "reduced" to economics or ecology or property (Fortes, 1969). But group rights to a common estate or lineage territory do seem to bear some regular relationship to land use and to influence the structure of the group itself. Although the regional band or tribe or community may claim and defend a demarcated territory, it is often a subgroup defined by kinship that corporately controls land. In effect, a discriminatory membership criterion, such as descent from common ancestor or kinship traced only through matrilineal links, is used to define exclusive rights to property. Tracing genealogical ancestry through one parent and assigning every individual unambiguously at birth to a discrete kin group allows certain close kin (e.g., sister's sons in a patrilineal system) to be excluded from rights to lineage land. Bilateral structures allow the individual choice as to what group he wishes to affiliate with and reside in. Under conditions of land shortage, a family group alone would have great difficulty in defending its property rights. A descent group organization can effectively exclude outsiders from land or, if circumstances allow, grant them land by adopting them as members or dependents (Gray, 1969:25).

As wealth in land or other durable property goes up with population density, settlement stability, and technological sophistication, the tendencies toward the regulated transmission of collective rights may give rise to unilineal kin groups (Forde, 1947:70). Among the horticultural Mae-Enga of highland New Guinea, Meggitt (1965) found that areas of land shortage were positively correlated with patrilineal organization or patrilocality. It appears that the descent groups were excluding affines (relatives by marriage) and nonkinsmen from access to land to ensure their own subsistence resources. Rappaport (1968:27-28) pointed out that a single New Guinea kin group with low population density might grant land rights to the abundant resource to a wide variety of relatives, but as the supply of open land declined and conflicts over farms and pigs increased, a tendency to confine use and inheritance to the more rigidly defined patrilineage would become apparent. A unilineal descent group can both reduce conflict for land among its members and secure cooperation beyond the nuclear family for the defense of scarce resources (Harner, 1970).

Population pressure does not, however, show a straight-line, isomorphic relationship to kin group estate formation. Indeed, further competition for the resource base eventually undermines the ability of the unilineal group to provide access to all its members, and such corporate groups then tend to disintegrate (Andrews, 1980). Comparing Maya-speaking populations of southern Mexico in Chan Kom, Zinacantan, and Chamula, Collier (1975) found that both sparse and very dense populations had little emphasis on living near and sharing land rights with patrilineal relatives, but the localized kin group was emphasized by those with an intermediate level of competition for resources (p. 206):

> Generally where land is abundant and a free good, swidden farmers do not have descent organization, but where land is scarce and a valued commodity, descent emerges to systematize right to land. When, however, land holdings are overly fractioned by inheritance, farming tends to give way to other occupations and land ceases to motivate descent-based kinship.

Under shifting cultivation, a relatively open community can accept various members with a variety of kinship links to a core of older or founding residents. As the fallow period shortens and distances to bush fields increase, groups of coresident, cooperating kin can compete successfully with more distant relatives or nonkin. The corporateness of the patrilineal descent group as a property-holding group is asserted. As in the case of the Ibo, when the land base becomes inadequate for the group and conflicting interests come to the fore, individual families withdraw to farm their own land or to seek employment without the obligation to share the proceeds with kin.

Tenure and the Market

Changes in tenure should also not be regarded as the mere product of internal population growth and efforts to maintain subsistence in a closed, circumscribed environment. Intensification today means in almost every case that not only is land being used more frequently and yields being increased or maintained but also the crops are often specialized for the market. Instead of producing a full range of subsistence crops, people are responding to demands generated outside their communities for food or condiments or fibers or animal products. Methods of production and often capital investments are dictated by the requirements for growing and processing the specific cash crop.

The Mandaya of Mindanao in the Philippines were traditionally shifting cultivators of hill rice swiddens in an area too wet to allow the burning of clearings (Yengoyan, 1971). Each year a family cleared a new field and built a house there. The family had rights to the crops and fruit trees on the current and old swiddens, and they had a claim to good locations with climax vegetation in the cross-country direction in which they were moving. Land disputes were rare, and because of low population density, more complex or confining tenure arrangements were not required. In the foothills, Mandaya are now engaged in the cultivation of abaca hemp. Hemp plants produce for 10-15 years, and the farmer needs stripping machines for reducing the leaves to fibers and draft animals for transporting the harvested plants to the house. Commercial production also requires larger fields and more labor from an extended family group that remains together rather than establishing separate neolocal residences, as in the hills.

Rights to land are still considered to result from usufruct, but the cultivator takes pains to keep his land planted in hemp or food crops. Claims are further enforced by fencing the land. Land is also given a money value. Though people feel that resources should be inherited equally, women, upon marriage, sell their inherited shares to their brothers, and in large families the land remains with the eldest son. To be productive, a hemp farm should not be fragmented and the capital goods, such as stripping machines and livestock herds, should be kept intact (Yengoyan, 1971:371). Since few suitable lands remain, the option of exploiting new areas is decreasing. Although inheritance rules are still amorphous, practices emphasize the maintenance of valuable continuing rights in hemp-producing lands and equipment by brothers with a multiple family household. Land tenure is being applied to the land itself rather than its crops, and inheritance is being concentrated in the patriline. Population pressure and agricultural intensification are hastened and their effects on land tenure reinforced by involvement in the production of hemp for the market.

Growing cash crops should not be regarded as automatically equivalent to the intensification of agriculture. Indeed, if land is readily available and the cheapest mode of production is temporary cultivation with minimal investment, there is little incentive to acquire more permanent rights to land. The same hill Kofyar who insisted on field boundaries marked by stones and took cases of theft and trespass to court did not acquire ownership of the migrant farms on the plains, where they grew yams, sorghum, and millet for sale. Because they were practicing shifting rather than intensive cultivation and they expected to move on every few years into untenanted forest land, they saw no need to establish title to their clearings. Instead they recognized the territorial political rights of the plains village chief by giving him an initial payment and a token share of the crops they produced in return for use rights to a patch of bush land (Netting, 1968:197). Arrangements were temporary, personalistic, and more in the nature of tribute than of rent. Scarcity of land along with permanent residence and agricultural intensification could be expected to regularize and systematize such loose tenurial arrangements, but the mere presence of a cash economy does not immediately evoke them.

Various Kinds of Coexisting Property Rights

In most technologically simple food-producing societies, there are a range of methods of varying intensity for utilizing different microenvironments in growing grains, vegetables, fruits, pasture and forage crops for livestock, and wood for fuel. If land use actually does affect land tenure, we might expect a corresponding spectrum of ownership rights. Even the general transition postulated by the 19th-century evolutionists from collective ownership of land by the clan to communistic household communities and eventually to individual family holdings never made all peasant land into private property. Engels (1972) realized that increasing population pressure and the lack of sufficient land to sustain shifting cultivation would have moved ancient German society toward more individual land tenure, but he also noted that communal lands persisted. "The arable and meadowlands which had hitherto been common were divided in the manner familiar to us, first temporarily and then permanently among the single households which were now coming into being, while forest, pasture land, and water remained common" (Engels, 1972:202). Resources that are needed by all but whose production is diffuse rather than concentrated, low or unpredictable in yield, and low in unit value tend to be kept as communal property with relatively equal, though not unrestricted, access by group members. Smaller, easily divisible, and more highly productive areas may be owned and inherited by individuals.

There are two contrasting subsistence strategies among the Bontoc Igorot of the northern Philippine mountains (C. B. Drucker, 1977). Terraced, laboriously irrigated, and intensively tilled wet rice fields provide the highly valued basis of the local diet. Sweet potatoes, black beans, and other cultigens are grown in unirrigated hill slope gardens with the length of fallow dependent on the distance of the garden from the village. Labor expenditure on the gardens is minimal and provided almost entirely by women. Tenure relates directly to the scarcity, value, and frequency of use of the land, and "the least restricted forms of inheritance operate upon the land holdings for which there is the least competition" (C. B. Drucker, 1977:7). Forests, pastures, and garden land pass from a single original claimant to all of his descendants, so that in time, a large bilateral descent group including virtually all village members has common rights of usufruct. If a rice terrace is constructed on such unimproved land, it becomes the private property of the builder. The difficulty of obtaining irrigation water means that few new terraces can be built, and most are acquired through a highly restricted inheritance system. Ideally, the eldest son inherits the rice fields of his father, and the mother's rice fields go to the eldest daughter. Extremely valuable heirlooms such as porcelain jars, gongs, and beads are inherited in the same manner (C. B. Drucker, 1977:9). Further inheritance rules specify the rights of illegitimate children, collateral descendants, younger children, and children by second marriages. The more valuable the property, the more restricted and detailed the system of inheritance along narrowly defined lines of descent. Poor people who garden extensively draw use rights from a wide range of bilateral kin, while the wealthy remember the genealogies that contain the inheritance histories of the rice fields they possess. Tenure rules and practices are appropriate to the scarcity value and economic importance of different types of land.

The coexistence of contrasting communal and individual rights to land in the same community for centuries can be documented for the Swiss alpine village of Törbel. The first written sources dating from the 13th and 14th centuries indicate that hay meadows, grain fields, vineyards, gardens, houses, barns, and granaries were owned by individuals who could rent, mortgage, or sell such properties (Netting, 1976). Bills of sale written by notaries and witnessed by fellow villagers assigned carefully demarcated plots, often with associated rights to irrigation water, to buyers for stipulated sums in cash. Partible inheritance, with each child receiving an equal share in the estate at the death or retirement of a parent, was the rule in Valais (Partsch, 1955) and is observed to this day (Netting and Elias, 1980). The attention devoted to the legal status and worth of such properties coincides with the scarcity of good agricultural land and water in the alps and with the early settlement and relatively high population of the mountain valleys. Land level enough to plow was in short supply, and the meadows had to be manured and watered frequently if the necessary two crops of hay were to be produced during the short summer. The two or three adult milk cows owned by the average family had to be sheltered and fed for eight months a year in costly log barns (Netting, 1972).

The proliferation of private property rights was paralleled, however, by old and persisting communal holdings. The high-altitude alp where all the village cattle were pastured together in the summer belonged to all village citizens, and, as early as 1483, outsiders were prohibited by charter from using it (Netting, 1976:139). Dividing it among private owners would have made access to the limited water sources and to the variable grazing locations more difficult, and would have called for a greater investment in fencing and more people to herd and milk the animals. In a similar way, every village family needed fuel for cooking and winter heating, but the cutting of the village forest had to be kept within its rate of regrowth, or a vital resource would have been destroyed. The forest also protected village lands from avalanches and prevented erosion on the steep slopes of the village watershed. Under these circumstances, the maintenance of communal rights was less a quaintly archaic custom than an institution for granting equitable access to an extensive resource that was needed by all and yet could not be used efficiently if it were divided.

Even more important was the management of natural, slow-growing plant resources whose overexploitation could directly threaten the village's capacity for survival. Rules limited the number of cows each owner could send to the alp to those he could feed from his own hay over the winter. Trees to be cut for firewood were marked each year by the elected village council, and the shares were divided by lot among the populace. Rights to communal holdings were strictly regulated. Only village citizens descended in the male line from ancestors living well before 1700 have been permitted access to common resources in the last several centuries (Netting, 1979). A democratically elected council and a village meeting of all adult males administered communal property and levied fines for any breaking of the rules. With membership in the commune jealously guarded and citizens scrutinizing every activity pertaining to the common property, the interests of both economically optimal use and conservation of a fragile resource base were served. The Swiss case suggests that the tragedy of the commons is not inevitable when users understand their environmental limitations and institutionally regulate the expression of their individual self-interests.

Where altitude, terrain, or water supply narrowly limit land use, we may predict that associated systems of land tenure will exhibit considerable stability. Where the same land serves several functions, especially in a complex stratified society, we can expect several layers of rights and duties as well as cyclic switches in the rules applied. The three-field system of land use prevailing in medieval western Europe and extending into the 18th century was based on a rotation of winter wheat, spring grains and legumes (the oats, peas, beans, and barley of the nursery song), and fallow (Bloch, 1966; Lynn White, 1962). The so-called common fields were divided into three large blocks, and individual landholders (villeins, tenants, or freemen) had scattered unfenced strips in each of the fields. Although the farmers may not have exercised ownership over the fields that belonged ultimately to a lord, a monastery, or other institution, they had rights to pass on the tenancy to their heirs. One of the symptoms of possible overpopulation in the early 14th century and of the desire of the landlord to prevent fragmentation of the dues-paying family holding was the licensing of heirs to marriage only when they received their fathers' land and the stipulation that noninheriting siblings could stay in the family only if they remained celibate (Homans, 1960). The common fields were also under partial control of the village community, which determined what was to be planted on the individual strips and when plowing, planting, and harvesting were to take place (Anderson, 1971). Once the grain was in, the livestock from the entire village was allowed to freely graze the stubble without regard for the boundaries of the individual strips. Thus proprietorship by a landlord, regulations of the agricultural cycle by a community, heritable usufruct of arable strips by a tenant, and grazing rights in common coexisted, and the temporary dominance of communal or individual rights was based on the phase of land use in the agricultural year.

A contemporary sample of 17 communities in the central Andes impressively supports the association of contrasting land tenure rules with differing agricultural regimes and ecological zones (Guillet, 1981). Such settlements typically use lands at different altitudes on the Andean slopes, having irrigated maize at lower levels, tubers grown by shifting cultivation at middle levels, and mountain grazing at high altitude. There are highly significant correlations ($P < .001$) between: (1) grazing areas and communal control, with indivisible use rights available to all members of the community; (2) tuber cultivation by sectorial fallowing, with divisible use rights extended to individuals and groups; and (3) continuous irrigated maize agriculture and specialized horticulture, with private control (Guillet, 1981:143). In the zone in which potatoes and other rainfall-dependent native tubers have traditionally been grown, a long fallow period must be maintained to allow for soil regeneration. Farmers coordinate their activities with rules on planting, harvesting, rotation of crops, and grazing. As in swidden cultivation, individual usufruct is compatible with the system, but communal control distributes necessery access to land and protects the fallow. "The lower the ecological zone, the more capable it is of sustaining intensification" (Guillet, 1981:147) and the stronger the tendency that it will be held as private property.

Land tenure in the central Andes is frequently seen in terms of a shift from communal to individual control due to the penetration of market forces that began in the colonial period. Models of both the diffusion of modern economics and of Third World dependence emphasize the external forces that transform common lands and ignore ecological constraints. Synchronic cultural ecology, by contrast, may underplay the impact of the market. Guillet (1981) convincingly argues for the influence of population pressure and market forces in the context of important ecological constraints at the local level that ultimately condition the form these pressures take with respect to land tenure.

Summary

The regularities emerging from basic research among both subsistence agriculturalists and market-oriented farmers in a wide variety of local environments, historic and contemporary, suggest that indigenous systems of land tenure do vary with land use. Shifting cultivation with long fallow periods, grazing on natural forage or crop stubble, and the cutting of woodlands are generally associated with communal territories. As population increases and resources become scarcer, group privileges may become more specific and exclusionary, and group membership may be narrowed and formalized through descent or other rules. More intensive land use, involving increased labor and capital investment, continuous production, permanent improvements, and greater competition for resources, moves tenure in the direction of more individualized rights with fixed boundaries, overlapping and hierarchical systems of ownership, complex inheritance rules, and provisions for transferring property rights by loan, lease, and finally sale. A market economy including agricultural products often fosters the specialization and more intensive land use that stimulate individual tenure, but unless the balance of population to resources changes decisively, there is little evidence that an internal (as opposed to an externally imposed) change in property rights will occur. As long as we remember that we are discussing gradations along a continuum rather than ideal, mutually exclusive types, we can break down land use into the constituent characteristics that vary along the spectrum of communal and individual tenure. Table 1 shows such a categorization. To the extent that these abstract characteristics of land use can be measured, they should correlate with observed changes in land tenure over time, such as in the Ibo and Mandaya cases, and with the range of coexisting property rights, such as in the Bontoc Igorot, alpine Swiss, and central Andean societies.

LAND TENURE AND EQUITY: THE CASE FOR THE MALADAPTIVE DISTRIBUTION OF PROPERTY RIGHTS

A frequent criticism of social anthropology and of ecological anthropology in particular is that, in their zeal to describe integrated, functional, and successful systems of human adjustment to the environment, the practitioners have overemphasized the continued vitality, equilibrium, and selective advantages of each cultural group. There is no doubt that anthropologists are concerned with how social structure and the economy work rather than why they fail. Ethnographies attempt to penetrate the strangeness and seeming irrationality of an alien culture to uncover the comprehensible meaning, sense, and symmetry. There is often an implicit contrast between the simple, satisfy-

TABLE 1 Relationships of Land Use to Land Tenure

	Land tenure type	
	Communal	Individual
Value of production per unit erea	Low	High
Frequency and dependability of use or yield	Low	High
Possibility of improvement or intensification	Low	High
Area required for effective use	Large	Small
Labor- and capital-investing groups	Large (voluntary or community)	Small (individual or family)

SOURCE: Netting (1976:144). Copyright © 1976 by Plenum Publishing Corp. Reprinted by permission.

ing life of a traditional community and the conflicts, poverty, oppression, and environmental destruction of modern industrial society. We must therefore be wary that our effort to understand the relationship of land tenure to land use does not lead us to take the position that property rights adjust themselves painlessly and with precision to changes in population, agricultural methods, and market conditions. The glaring and often growing inequality in the distribution of land and the international volatility of the land reform issue convince us that this is not the case. Why is it, then, that the obvious benefits of enjoying the fruit of the tree one has planted or willing to a child a carefully tended rice paddy can logically lead to the absentee ownership of great estates worked by landless laborers? Are individual property rights a key to the greatest good for the greatest number or a tool for the institutionalization of greed and rapacity?

It is apparent that the same or very similar rules of individualized tenure may be applied in a community of self-sufficient, land-owning peasants, such as the mountain Swiss, and, on a large estate, such as that held for centuries by the Medici family in northern Italy (McArdle, 1978). Examples with some historical time depth allow us to see the unequal distribution of property actually taking place. There is no reason to think the process is a new one. As soon as denser, sedentary populations and pressure on the most desirable land and water supplies became evident, some families, because of differential demographic growth, judicious marriages, or political coercion could secure a holding larger than average. By protoliterate times in Mesopotamia, there were already recorded differences in the ownership of land, and purchase prices varied widely, "suggesting important differences in productivity, ease of access, assurance of adequate irrigation water, devotion to intensive orchard cultivation, or the like" (R. McC. Adams, 1966:55-6). Where corporate landholding groups were mentioned in Early Dynastic deeds, one or a few individuals are listed as owners of the field or recipients of its price in barley, silver, or other commodities. Others called "sons of the field" or "brothers of the owners" seemingly gave their assent to the sale in return for a small gift of food (R. McC. Adams, 1966:83-84). It is possible that the group involved was a corporate kin group or "conical clan" already distinguished as richer landholders and relatives with only nominal rights to the property.

Irrigation is a prime example of agricultural intensification, and it can be seen as a capital improvement related to population density and produced by labor that adds to the value of land (Dumond, 1972). The practice of shallow well pot irrigation in prehistoric Oaxaca, the expansion of population beyond the high-water-table zone, and competition for the limited highly productive land may have led to "initial disparities in wealth and status" (Flannery, 1967). Canal irrigation represents heavier investment to create valuable land and may further restrict access to land. Flannery (1967) reports that there tends to be a less equitable distribution of land and property rights in contemporary canal irrigation communities than in pot-irrigating communities. If investment in the building or repair of an irrigation facility is beyond the capacity of a single community, a wealthy outsider may exchange assistance for rights to revenues from village lands, charges for water use, or an irrigated estate. When an underground channel or qanat serving the Iranian village of Deh Salm was blocked by silt, a provincial dignitary was persuaded to open it only by being awarded several days of water flow, a share that had considerable market value and could be used to irrigate new land (Spooner, 1974). With competition for land and water and with few alternate subsistence resources available, a minor advantage in location, in inheritance, or in agricultural luck during a bad year can be translated by loans, sales, or investment into an expanding share of local land or more exclusive rights.

Land Consolidation and Creation of a Landless Class in European History

The process of concentrating the control of landed property in the hands of a few while the average farmer has his homestead reduced in size or eliminated has been repeated at least three times on a massive scale in Europe. Growth in rural population produced symptoms of agrarian distress in the periods 1250-1350, 1550-1650, and 1750-1850 that are also visible in the underdeveloped world today (Grigg, 1980). As increasing numbers of cultivators competed for a fixed supply of land, farms were subdivided until they became too small to provide a subsistence living for a family or provide the holders with full-time employment. Landlessness grew both absolutely and as a proportion of the rural population. Food prices rose as demand went up faster than supply, and the same forces drove rents and the prices of land upward. Unemployed farm laborers and peasants seeking additional employment competed for jobs causing a fall in real wages. At the same time that rural poverty increased, landlords and large farmers found low wages and rising food prices profitable, allowing them to buy up or "engross" the properties of their less fortunate neighbors (Grigg, 1980: 286-87). Inflation benefited the rich capital accumulators while the poor were squeezed (Wallerstein, 1977).

Among the French peasants of Languedoc who survived the black death of the 14th century, land was abundant and available for the taking or at favorable rates. The shortage of labor led to high wages, and the many families with medium-sized farms had a substantial and varied diet (LeRoy Ladurie, 1974:44, 49, 87). With population growth, there was "a veritable pulverization of rural property" in the 16th century as large landholdings increased, small ones proliferated, and those of medium size almost disappeared (p. 21). "Microproprietors," who had inherited too little land to support themselves, had to purchase grain for subsistence and accept seasonal day labor on the big estates. As the gap widened between population and production and between prices and wages, rural capitalists made loans in grain or money and in time took over the fields of their debt-ridden clients (p. 126). Marginal land was taken out of pasture and put into low-yielding grain fields, further diminishing meat supplies and the manure required for maintaining soil fertility. In much of Europe harvest failures in the 15th century further impoverished the population and sent armies of uprooted paupers onto the roads (p. 137).

The study of social, demographic, and economic processes at work over time in a region or a few villages has been pioneered by French social historians and geographers of the Annales school, and it can with justice be termed an anthropology of history. The changing distribution of land is placed in a comprehensive ecological framework, and controlled comparisons of differing adaptive responses allow us to go beyond stereotyped, single-cause analyses of tenure types. For instance, the equilibrium of several English villages in the Forest of Arden was disturbed in the latter part of the 16th century by sharp population increases due to relatively high fertility, low mortality, and high immigration (Skipp, 1978:65). Many of the newly formed households lacked access to adequate landholdings, forcing them to carve poor farms out of the common waste-lands or rent cottages from larger landowners. Positive responses to the resulting situation of low wages and spiraling food costs were increased employment in cottage industry and crafts such as weaving, nail making, and carpentry. The prosperous larger farmers also increased production of grains, dairy products, and wool by enclosing pastures, installing new systems of crop rotation, ditching and draining their fields, and employing more agricultural equipment and draft animals (Skipp, 1978:65). But occupational specialization and more intensive cultivation were not sufficient to protect the growing landless population (up from 5 percent to 32 percent), whose lack of a dependable subsistence base rendered them vulnerable to any economic downturn. Bad harvests or declines in the regional or international market for their goods left many people destitute, causing a drop in births and an increase in infant and maternal mortality that was not experienced by the landed class. Various local charities and the parish relief mandated by the Elizabethan Poor Laws could provide only minimal help to the landless third of a polarized community (Skipp, 1978:85-87).

The redistribution of land and the more marked stratification of the rural population I have been discussing were not usually accompanied by changes in the laws of land tenure. Whether a landlord met population growth by subdividing his estate into progressively more tiny and less secure tenancies, as in Ireland between 1750 and 1848, or whether he evicted tenants and converted arable holdings to sheep, as in 16th-century England, depended not on the statutes but on such factors as the prices of grain and wool, or the capacity of the potato to provide cheap food from small plots (Grigg, 1980:119, 88). Common lands did not suddenly vanish overnight but were instead whittled away by squatters needing a cottage site and a garden or by a lord appropriating part of the common grazing for his own sheep (p. 70, 90). Villagers who depended on grazing rights for their livestock, the use of woodland for fuel and timber, and rushes and fish from the marshes stoutly resisted incursions. As pressure on the commons increased, rules for "stinting" the number of cattle allowed, fixing boundaries, and limiting rights to village landholders were drawn up and enforced. The poor, who depended mone on common pasturage and the products of gathering, were particularly hurt by such restrictions (Jones, 1974; Spufford, 1974).

Legal change, when it came, may have been less decisive than is generally believed. The English parliamentary enclosure of common fields in the 18th and 19th centuries reduced fragmentation but did not rapidly displace small landholders in favor of large farms. The process of consolidation had been proceeding gradually long before the enclosure acts (Grigg, 1980:170). Rural exploitation and misery have often been attributed to the political and military coercion of rich landlords and governmental leaders (Grigg, 1980:292).

> Contemporaries believed that inequitable land-tenure systems as in Ireland, or the Poor Laws, as in England, were the main cause of poverty. But the universality of the phenomenon, in areas of quite different land-tenure systems, suggests that the rural population had grown too large to provide farms of adequate size for all the rural population, or employment for all the landless.

While the laws and customs regulating land tenure give the impression of great stability and historic continuity, the actual distribution of land and the uses to which it is put are responses to variations in population, employment opportunities, prices, wages, world trade, and agricultural technology.

Third World Land Tenure and Land Reform

The habit of seeing land tenure as a prime mover in agricultural development or as a sword to cut through generations of colonialist or class domination in the Third World is still very much with us. Few issues so polarize the political right and left as does land reform and policies aimed to secure equality and economic growth. One view is the belief that technologically simpler societies are close to a past in which land was held in common by the group and there were few distinctions of property or wealth within the society. This heritage can supposedly be modernized by adapting it to group or state farms, cooperatives, or rural communes. Such institutions not only avoid capitalist exploitation and the gross inequality of kulaks and plantation hands but also provide for saving and investment, give economies of scale, and allow the efficient introduction of machinery. A contrary point of view sees private property as a prime achievement of evolutionary progress in which personal profit in a system of free enterprise best motivates the hard work, financial investment, planning, and innovation necessary to modern cash-crop farming. The same irrefutable economic logic is seen to apply to the family farm and the corporate agribusiness, the competition of the marketplace infallibly picking the winning enterprise.

Basic research casts doubts on the ideological certainties and preconceptions that imbue so many discussions of land tenure. With cross-cultural examples of the permutations of territoriality, usufruct, coexisting but contrasting property rights, and historic changes in land distribution, it also becomes necessary to approach plans for land reform in a cautious and somewhat clinical manner. We may be safe in saying, however, that the effort to change land tenure alone without attention to patterns of land use, population, and other variables in the ecosystem is a cart-before-the-horse reversal of priorities.

The recent situation of political struggle, civil war, and international confrontation in El Salvador highlights relevant issues with an urgency often missing in academic considerations of land tenure. We are fortunate that social scientists have dealt with both theoretical and practical aspects of this question. The same pattern of population growth, unemployment, landlessness, and consolidation seen in the historic European cases is visible in El Salvador. Population increase has followed the familiar path of exponential growth, reaching an annual rate of 3.49 percent a year for the period 1961-1971 (Durham, 1979:22-3). Beginning in the mid-1950s, the food supply failed to keep up (p. 22), and the population has shown a growing incidence of poverty and malnutrition along with the highest rate of homicide in the world (Chapin, 1980). The density of more than 200 people per square kilometer is the highest in Latin America. Subsistence crop cultivation has been extended to marginal hill areas, leading to deforestation and declining yields (Durham, 1979:80). The limited agricultural resources are increasingly unequally distributed. The percentage of rural landless people has grown from 11.8 in 1950 to 40.9 in 1975 (Chapin, 1980). While 125 estates are more than 1,000 hectares in size, 71 percent of the farms have less than 1 hectare each. Access to land was directly related to the survival of children in one survey, with 48 percent mortality among landless families declining to about 20 percent for families owning 2.5 hectares or more. Tenants had appreciably higher child death rates than mid landowning families (Durham, 1979:85-88, 91).

Although El Salvador appears at first glance to be a straightforward if frightening case of Malthusian positive checks about to be imposed, closer inspection shows that food and land are not absolutely in short supply. Total agricultural production has kept pace with population, but since 1955 land has been taken from growing the food crops of maize, beans, rice, and sorghum and put into the export crops of coffee, cotton, and sugar (Durham, 1979:31). Encroachment on Indian lands by haciendas had begun shortly after Spanish conquest in the 16th century, but the communal territories used for shifting cultivation were abolished by decree in 1881: "The existence of lands under the ownership of communidades impedes agricultural development, obstructs the circulation of wealth, and weakens family bonds and the independence of the individual. Their existence is contrary to the economic and social principles that the Republic has accepted" (quoted in Durham, 1979:42). The legal move in this case did not reflect changing land use of the majority of small farmers but furthered the interests of coffee planters attempting to acquire privately owned tracts. Dispossessed peasants became hacienda laborers, and later dips in world coffee prices along with the Great Depression led to further consolidation of properties. The loss of land led to rural workers' uprisings and violent repression (Durham, 1979:44). Both changes in tenure and land distribution were not gradual and cumulative as in the European cases, but imposed from above by a dominant mercantile minority. Durham (1979:54) credits food scarcity less to population growth than to the large areas underutilized (as in grazing lands) or devoted to export crops. "Land is scarce not because there is too little to go around, but rather because of a process of competitive exclusion by which the small farmers have been increasingly squeezed off the land--a process due as much to dynamics of land concentration as to population pressure" (Durham, 1979:54).

The earlier history of El Salvador, like that of other Latin American countries, exemplifies a conscious political campaign, first by the colonizers and then by the national elite, to abolish communal jurisdiction over land. A major defense of the Indian closed corporate community was its claim to immemorial control of common lands that were periodically reallocated to members but could not be transferred to outsiders (Wolf, 1957). Well before population growth forced changes in land use, "land was to become an object to be bought, sold, and used, not according to the common understandings of community-oriented groups, but according to the interests of nation-oriented groups outside the community" (Wolf, 1955).

With abundant labor and low wages, the traditional Salvadorean arrangement of giving resident hacienda laborers small wages and a subsistence plot has broken down. Laborers must now by law be paid wages, and their plots are rented. For seasonal tasks, short-term workers are hired, and these landless people live in rural shanty towns. Hacienda owners are often absentee, living in the capital or abroad. Land reform under these circumstances of poverty and greatly unequal access to crop lands was an obvious necessity. The so-called Land to the Tiller program, developed with American advice, paid little attention to land use or to practical means of implementation. A new law announced over the radio entitled all renters to take over land they were now renting, with compensation to be paid to owners over a period of years (Chapin, 1980). Since this land was predominantly hill land planted in subsistence crops farmed for a single year by shifting methods, it would have little value to the renter as a permanent possession. Such plots were characteristically too small to allow for fallow periods with continued slash-and-burn cultivation. The renters themselves objected to taking over the land of small landholders, the elderly, and widows. Legal rules for obtaining title were unclear, boundaries were often in dispute, and the expense of litigation and lawyers in the city was more than most rural laborers could support. The goal of private property for the poor could thus be subverted by the officials and the courts whose duty it is to guarantee such rights. Resident owners of medium-sized farms were often hostile to renters, summoning the national guard to force the renters off the land and refusing to return their payments (Chapin, 1980). In such cases, the renter was fundamentally powerless. That temporary usufruct of a tiny swidden was a good basis for establishment of a permanent, self-sufficient farm is a ludicrous contention. Nevertheless, an American law professor who had helped to institute the land reform claimed that tenancy had been eliminated. "Former tenants farm plots of about two acres. But as owner-operators they will be motivated to exploit their family farms more effectively, as shown by the land reforms in Japan, Taiwan and South Korea with even smaller plots" (Prosterman, 1981).

In another phase of the agrarian reform, estates of more than 500 hectares were taken over by the government.

Old employees were to be kept on, workers' councils established, and administration provided by government technicians. The maintenance of these larger land units was deemed necessary to the continued production of cash crops on which the Salvadorean economy depends. Though the plan appears to have popular support, there have been difficulties in promptly obtaining credit, seed, fertilizer, gasoline, and other necessary capital inputs through the state bureaucracy (Chapin, 1980). Moreover, the workers who receive the same pay as before and have in most cases no effective participation in the direction of the enterprise that technically belongs to them feel that they have only exchanged one boss for another. Their insecurity is heightened by threats of armed attacks from both the political right and the left and by the sense that the hacienda owners may yet return to reclaim their property and punish the workers.

Wage laborers on the estates continue to work in organized groups on cash crops, but they want individual subsistence plots to ensure their food supplies. Such grants of land were not included in the legislation that envisioned group farming for family subsistence grains. Changes in land use might also benefit workers. Growing only one crop, such as sugar cane or cotton, tends to concentrate work seasonally with intervening periods of underemployment and few alternate jobs (see Yengoyan, 1974). The few diversified estates in El Salvador produce cacao, coffee, citrus, corn, salt, and pond-raised fish. This provides year-round work for a larger labor force and allows the support of a school and a cooperative store on the estate (Chapin, 1980). The economic benefits of large, market-oriented, collective farms are not inconsistent with the private control of subsistence plots by workers, increased worker participation in administration, and more efficient land use through agricultural diversification. Haciendas will certainly differ in the degree to which collective or individual land tenure is viable according to the types of crops grown (Chapin, 1980). Land reform of any sort, based as it must be on understanding the legal changes and widespread voluntary compliance, has little possibility of success in the midst of civil disorganization, fear, and violence.

THE MEANING OF LAND AND THE IDEOLOGY OF CONFLICT

In pursuing ideas about how territoriality and land tenure work, about how the rules and practices interact with such objective variables as population, land use, and political power, it is easy to lose sight of the way that people think about land. An ethnoscientific approach to land tenure would have emphasized the words and classifications that structure shared systems of concepts about how resources may be held and how rights are acquired. The use of native terms, such as the Tiv word tar, which denotes a neighborhood of compounds and farms whose adult male residents tend to be members of the same patrilineage, conveys a sense of the kin group temporarily occupying portions of the landscape. Tar is not easily translated into another language (P. Bohannan, 1963:103):

> Every Tiv has a right to an adequate farm on the earth which holds his tar. This is a right to a farm, not a specific piece of land. A farm lasts only for two or three years, then reverts to fallow and the specific right lapses. However, the right to some farm in the tar never lapses. Thus, the position of a man's farm varies from one season to the next, but his juxtaposition with his agnatic kinsmen, and his rights to a farm, do not change.

The Tiv use spatial distance between households as an index of genealogical relatedness, and their migrations as shifting cultivators across the Nigerian savannah may lead them to reorient their lineage ties to agree with their territorial positions (L. Bohannan, 1952; P Bohannan, 1954). The continually adjusting Tiv genealogical map is a radically different conception from the Western practice of dividing the earth's surface by use of a rigid imaginary grid based on precise positions of the stars and exact survey measurements (P. Bohannan, 1963). The primacy of the social group and the impermanence of the bush-fallow fields is further emphasized by the Tiv claim that "We have no boundaries, only arguments." If we are correct in connecting descent group territorial rights and individual usufruct to a type of agriculture, the Tiv cognitive cartography fits neatly with the requirements of their subsistence system. Our ethnocentric notions that land must be a measurable entity divided into thing-like parcels and that a person can have rights to a piece of the map are appropriate to a different type of land use, with its accompanying mathematical and technical processes.

Anthropologists have long noted the tendency of every human group to exhaustively name the places in their habitat that they use and to transmit a catalogue of its natural features in their conversation. The territories of Australian aborigines are threaded by the detailed journeys, campsites, and magical encounters of their ancestors. The myths that recount such travels represent a symbolic claim to a particular area and may provide a traditional guide encoding the distribution of resources (Peterson, 1979). The place names that peasants give to every field, valley, grove, and stream are an oral map of their landscape, referring not only to distinguishing features and agricultural activities but also to the history and past owners of fields. The smaller and more precise the named divisions, the greater the nearness to human habitation, frequency of use, and value of products.

Actual tenure and use are bound up in a continual dialogue with the culturally created and transmitted meanings of property and work. The ranchers who settled southern Saskatchewan in the late 19th century sought out pasture, water, and shelter for their cattle, moving the animals about on the open range to take advantage of variable resources in a semiarid environment (Bennett, 1969). Theirs was an ideal of using the natural environment without substantially altering it, conserving the wilderness, and owning stock rather than land. Except for the sites of ranch buildings, much of the grazing land even today is government-owned and used under long-term, heritable leases by the ranchers. Like the alpine pastures of the Swiss, such marginal, low-yielding resources may be best administered under some form of ultimate communal control. The widely dispersed settlement and self-reliance of ranching encouraged the same attitudes of individualism, independence, and hospitality that characterize other pastoral peoples (Goldschmidt, 1971). Ranchers were originally pushed into the hills by farmers taking up homesteads on the more fertile plains. The plan of allotting 160 to 320 acres of land to anyone who would settle and cultivate it was designed by the Canadian government on the model of what one man and a team of horses could cultivate in the eastern part of North America. This use of limited, privately owned land was ill-adapted to the dry-land, large-scale growing of wheat that was one of the few viable agricultural systems in western Canada. Farms developed on a rectilinear grid without regard to the variability of local resources often failed in the early years, and the remaining enterprises had to expand and invest heavily in machinery. Farmers adopted the ecological posture of subduing and manipulating the land, fencing it, irrigating, and planting greenery to make their homes resemble those of more humid areas (Bennett, 1969:85-94). Farmers also organized grazing co-ops in order to raise cattle, and they entered politics to secure crop subsidies and disaster insurance from the government. Both ranchers and farmers have been forced to accommodate fixed legal systems of land tenure to the exigencies of quite different kinds of land use and to correspondingly different conceptions of the natural environment.

Though Western ideas of rights to land may collide obtusely with the practical mental maps of group territories held by hunter-gatherers, herders, or shifting cultivators, an even more relevant issue for basic research is the manner in which conflict over unequal systems of land distribution is disguised or hidden by public disputes over politics, morality, and ideology. The theme of much of this paper has been the rationality and resourcefulness that societies bring to the task of adapting property rights to subsistence means, environmental conditions, and economic changes. When this process has been allowed to develop from within rather than being imposed from above, it has often meant a very gradual evolution of the rules along with rapid adjustment of usufruct practices, inheritance, and variant forms of tenure. Except in some cases of revolution, when slogans of private property or communistic control of land may be bandied about, most societies most of the time appear to agree on the most general cultural premises of ownership. Yet individuals are keenly aware that, while playing under the accepted rules, their shares of necessary resources may be less than they need and deserve. It is striking, however, that major public conflict is seldom phrased in terms of the distribution of land.

An example of different underlying relationships to the means of production is seen in the case of two factions of Sikh peasant cultivators in an Indian village of the Punjab (Leaf, 1973). A statistical analysis indicates that the groups contrast with one another in (1) the amount of land per household, (2) crop income per capita, (3) percentage of landless members, (4) frequency of tenancy and sharecropping, and (5) the use of mortgaged land rather than the more secure owned land. The opposition is not between landlords and tenants but between relatively land-rich and land-poor factions. Although conflicting interests were expressed directly in court litigation over ownership of land, with each group supporting its own members as witnesses, most of the factional competition was conceptualized by the participants as based on traditional issues of religious and moral values, alliance with national political parties, and protection of the welfare and property of one's family. Leaf (1973) argues that the real material differences in access to land are obscured by the welter of ecological and economic factors influencing individual decisions on land and labor use. On the other hand, a formal system of conventionalized concepts allows simple dichotomies between religious and irreligious behavior, political party supporters and rivals, and kin aligned against nonrelated enemies. Lacking the simplifying (and limiting) tools of social science, the Indian peasants generalize on their experience, using distinctive, binary, moral, and political attitudes to create a uniform and relatively stable framework of socially meaningful opposition. The ideological split of factions is a working metaphor for genuine, if unperceived, conflicts of interest in rights to land.

It is not realistic to treat social conflict as in large measure an unconscious translation of competition for land and resources, yet it is also inadequate to compartmentalize religious, political, and kin-based rationales for opposition as if they had little to do with each other or with land tenure and use. In fact, the inquiries of social science into ideological confrontations take on a new dimension and persuasiveness when set in an ecological context. Where rules of land allocation are glaringly at odds with supply and demand for resources, conflict may bring about adaptive change, but the fighting is seldom explicitly over property. The Puritan settlements established in the Massachusetts Bay Colony wilderness included citizens from open field and from enclosed farmstead English villages (Powell, 1963). Men of substance, clergymen, and elected officers of the new communities were granted larger farms and more favorable house sites than their neighbors, but the principle of private property and inheritance was brought intact from the homeland. When further immigration and natural growth began to strain local resources, there were quarrels over rights to common meadowland and over the inequality of the original shares. But the dissidents who moved away to found new, independent towns on the frontier appealed always to religious justifications. They cited doctrinal differences and the requirement to worship as they pleased under their own minister. Sectarian controversy was the language in which squabbles over land were ultimately stated and resolved.

In the same manner, the poor farmers of the Salem village hinterland campaigned for years to have an independent church and an appropriately supported minister. Two generations of offspring subdivided the farms, and the town border prevented expansion. An average Salem village landholding declined from 250 acres in 1660 to 124 acres in 1690, and there were sharp boundary disputes with neighboring towns (Boyer and Nissenbaum, 1974:90-91). When the small landholders' envy and frustration against the more affluent merchants, innkeepers, and gentlemen farmers near Salem town boiled over, it took the bizarre form of witchcraft accusations directed by the children of back country farmers against the wives and retainers of their more prosperous neighbors (Boyer and Nissenbaum, 1974). The quality of land and access to the market polarized the community, but the factional split gained its bitterness and destructiveness from the moral fervor and the psychological substratum of magicoreligious belief with which it was acted out. Rights to land alone without the trappings of good and evil or true-believer politics seldom exercise such emotional power over the holders. Land tenure, as the rules and practices of competition for scarce and vital means of subsistence, will perhaps always appear freighted with fear, suspicion, and a popular ideology of conflicting social forces. It refuses to be confined to the realm of practical reason and objective judgment.

SOME CONCLUDING COMMENTS

What anthropologists have learned about territoriality and land tenure has often come about as a by-product of research aimed at other goals, such as understanding foraging strategies or kin group structure or agricultural intensification. There are remarkably few general theories or pronouncements on the cross-cultural regularities of property holding in the literature. Social scientists have usually been content to show how rights to resources were allocated in particular societies. I have not even bothered to confront the persistent popular stereotypes of primitive communism or evolutionary "progress" toward private property or to indicate how inadequate they are in understanding our own case studies.

By not focusing on property per se, anthropologists have succeeded in seeing behavior and beliefs concerning ownership in the context of larger economic and social systems. While politics and law often isolate land tenure from other cultural factors, recent anthropology has insisted that ecological relationships, including the physical environment, agricultural uses, labor expenditure, technology, market forces, and historic ideologies, must all be considered. The field work tradition of long-term participant observation has gone beyond the often ethnocentric questions and misleading answers to the query, "Who owns what?" The patient compiling of cases in which individuals use resources, groups claim territories, and people inherit, loan, rent, give, sell, and argue over property exhibits the complexity and variation of culturally recognized rights. Rules or statutes are important bits of data, but they by no means exhaust the meanings and permutations of property in a single society. Nor is a change in the law by itself likely to change the agricultural system or profoundly alter the distribution of wealth. Property rights, especially in self-sufficient subsistence societies, are apt to reflect closely the ecological requirements for spacing populations, adjusting to demographic changes, and preserving a working balance with the environment. Hunter-gatherer territoriality exemplifies this sensitive and pragmatic accommodation.

As resources become more localized and dependable and as competition among settled groups for access to these resources increases, groups become permanently attached to the sites of their major subsistence activities. Fishermen may claim and defend choice areas for harvesting salmon or trapping lobsters, and rights to use marine territories become possessions to be transferred in public potlatches or by recognized inheritance. Scarce land becomes a demarcated, guarded estate with access limited to those entitled by membership in a descent group or a corporate village. Communal tenure allies individuals to prohibit resource use by outsiders and equitably allocate usufruct rights among members. Such measures serve also to limit destructive overexploitation of the environment and promote conservation.

Adaptation of land or water rights to land or water use can be postulated but not demonstrated when the basis is merely a brief acquaintance with another society. When the system changes, however--when population grows or a new technology is adopted or the market economy expands--then alterations in land tenure can be related causally to other variables. Restudies or comparisons or local histories can show us these processes at work. We can see how hunter-gatherer territoriality is modified by the occurrence of water supplies or how the replacement of shifting cultivation by permanent irrigated fields modifies communal tenure in the direction of individual rights to land. In similar fashion, cash crops like cacao or hemp may require special soils, investment, and long-term production that increase the value of land and make defined rights, marked boundaries, and rules of inheritance more likely. The same community can allow all its members access to low-yielding, unpredictable grazing land and forests while assigning grain fields, barns, and vineyards to individual ownership. Such cases demand an attempt to explain land tenure in terms of land use and to incorporate rights to resources in models of ecological systems. They also express a confidence in the ability of local groups to perceive and institutionalize rational means for allocating scarce goods and regulating internal competition. Community regulation can avert even the tragedy of the commons.

We cannot, however, adopt some rosy view of societies unerringly acting to produce the greatest good for the greatest number. When they are no longer isolated and subsistence-oriented, the same processes that led to increasing individualization of tenure may result in great inequality in the distribution of resources. Population growth in a money economy with high food and land prices and low wages may further consolidation of large properties and the growth of a sector of landless laborers. Wealth and misery can result with no change in the rules of landholding. Concepts of private property that evolved under circumstances of population pressure and relative resource scarcity can also be transferred to inappropriate contexts through conquest, colonial domination, or commercial exploitation. Inequities grow not from within a system but by imposition and coercion from outside. Laws and economic arrangements become instruments to preserve and extend a distribution of property that benefits only an elite. Land tenure in such circumstances may thwart effective land use and the optimal application of labor.

Land reform, an obvious necessity in much of the Third World, should not merely overturn a system on doctrinaire political grounds. It must make policy with the aim of suiting tenure to use and allowing for the flexibility and multiple options by which local groups can meet their special agrarian needs. Centrally conceived, rigidly applied blueprints of land reform may be simple, quickly installed, and bureaucratically satisfying, but they leave out the real experts, the local farmers whose practical insights are untapped and whose ability to organize for common goals is ignored. Basic research has suggested some principles of land tenure that appear to have cross-cultural validity. The value of an ecological approach to land reform policy is that it emphasizes the specific context of environmental variables, population, agricultural systems, market factors, legal standards, and political power as well as the particular history of a local ecosystem. The hypotheses we have developed, tentative and middle-level though they are, point to the most significant areas for data collection--they tell us what we need to find out. Our firsthand experience through anthropological field resesrch alco leads us to value the decisions people make about property and the institutions they create to cope with conflict over resources. The people we live with and learn from can show us how their system of land tenure works and, often, how it can change most beneficially.

REFERENCES

Acheson, James M.

1975 "The lobster fiefs: economic and ecological effects of territoriality in the Maine lobster industry." Human Ecology 3:183-207.

Adams, J. W.

1973 The Gitksan Potlatch: Population Flux, Resource Ownership and Reciprocity. Toronto: Holt, Rinehart and Winston of Canada.

1981 "Recent ethnology of the Northwest coast." Annual Reviews in Anthropology, Volume 10.

Adams, Robert McC.

1966 The Evolution of Urban Society: Early Mesopotamia and Prehispanic Mexico. Chicago: Aldine.

Anderson, Robert T.

1971 Traditional Europe: A Study in Anthropology and History. Belmont, Calif.: Wadsworth.

Andrews, Tracy

1980 "Agriculture, descent and land tenure: an anthropological perspective." Unpublished preliminary examination paper, University of Arizona.

Barnard, Alan

1979 "Kalahari Bushman settlement patterns." Pp. 131-144 in D. C. Burnham and R. F. Ellen, eds., Social and Ecological Systems. New York: Academic.

Benedict, Ruth
1934 Patterns of Culture. New York: New American Library.

Bennett, John W.
1969 Northern Plainsmen: Adaptive Strategy and Agrarian Life. Chicago: Aldine.

Bishop, C. A.
1970 "The emergence of hunting territories among the Northern Ojibwa." Ethnology 9:1-15.

Bloch, Marc
1966 French Rural History. Berkeley, Calif.: University of California Press.

Boas, Franz
1921 Ethnology of the Kwakiutl. Bureau of American Ethnology, 35th Annual Report.

Bohannan, Laura
1952 "A genealogical charter." Africa 22:301-315.

Bohannan, Paul
1954 "The migration and expansion of the Tiv." Africa 24:2-16.
1963 "'Land,' 'Tenure' and Land-Tenure." Pp. 101-111 in D. Biebuyck, ed., African Agrarian Systems. London: Oxford University Press.

Boserup, Ester
1965 The Conditions of Agricultural Growth. Chicago: Aldine.

Boyer, Paul, and Stephen Nissenbaum
1974 Salem Possessed: The Social Origins of Witchcraft. Cambridge, Mass.: Harvard University Press.

Bronson, B.
1972 "Farm labor and the evolution of food production." Pp. 190-218 in B. Spooner, ed., Population Growth: Anthropological Implications. Cambridge, Mass.: MIT Press.

Brown, Paula, and Aaron Podolefsky
1976 "Population density, agricultural intensity, land tenure and group size in the New Guinea Highlands." Ethnology 15:211-238.

Chapin, Mac
1980 A Few Comments on Land Tenure and the Course of Agrarian Reform in El Salvador. Washington, D.C.: Agency for International Development.

Clapham, W. B.
1973 Natural Ecosystems. New York: Macmillan.

Clarke, W. C.
1966 "From extensive to intensive shifting cultivation: a succession from New Guinea." Ethnology 5:347-359.

Codere, H.
1950 Fighting with Property. New York: Augustin.

Collier, George A.
1975 Fields of the Tzotzil. Austin, Tex.: University of Texas Press.

Conklin, H. C.
1957 Hanunoo Agriculture. Rome: FAO.
1961 "The study of shifting cultivation." Current Anthropology 2:27-61.

Cowgill, G. L.
1975 "On causes and consequences of ancient and modern population changes." American Anthropologist 77:505-525.

de Janvry, Alain
1980 The Role of Land Reform in Economic Development: Policies and Politics. Paper presented at the annual meeting of the Allied Social Science Association, Denver, Colo., September 6.

Demsetz, Harold
1967 "Toward a theory of property right." American Economic Review 57:347-373.

Donald, L., and D. H. Mitchell
1975 "Some correlates of local group rank among the Southern Kwakiutl." Ethnology 14:325-346.

Drucker, C. B.
1977 "To inherit the land: descent and decision in Northern Luzon." Ethnology 16:1-20.

Drucker, P.
1939 "Land, wealth, and kinship in Northwest Coast society." American Anthropologist 41:55-64.

Drucker, P., and R. F. Heizer
1967 To Make My Name Good. Berkeley, Calif.: University of California Press.

Dumond, D. E.
1961 "Swidden agriculture and the rise of Maya civilization." Southwestern Journal of Anthropology 17:301-316.
1972 "Population growth and political centralization." Pp. 286-310 in B. Spooner, ed., Population Growth: Anthropological Implications. Cambridge, Mass.: MIT Press.

Durham, William H.
1979 Scarcity and Survival in Central America:

Ecological Origins of the Soccer War. Stanford, Calif.: Stanford University Press.

Engels, F.
1972 The Origin of the Family, Private Property, and the State. Originally published in 1884. New York: International.

Flannery, Kent V., A. V. T. Kirkby, M. J. Kirkby, and A. W. Williams
1967 "Farming systems and political growth in Oaxaca." Science 158:445-454.

Forde, C. D.
1947 "The anthropological approach in social science." British Association for the Advancement of Science 4(15):213-224.

Fortes, Meyer
1969 Kinship and the Social Order. Chicago: Aldine.

Freeman, J. D.
1955 "Iban agriculture." Colonial Research Studies 18. London: Colonial Office.

Geertz, Clifford
1963 Agricultural Involution. Berkeley, Calif.: University of California Press.

Gleave, M. B,. and H. P. White
1969 "Population density and agricultural systems in West Africa." Pp. 273-300 in M. F. Thomas and G. W. Whittington, eds., Environment and Land Use in Africa. London: Methuen.

Goldschmidt, Walter
1971 "Independence as an element in pastoral social systems." Anthropological Quarterly 44:132.

Gray, Robert F.
1969 "Introduction." In R. F. Gray and D. H. Gulliver, eds., The Family Estate in Africa: Studies in the Role of Property in Family Structure and Lineage Continuity. London: Routledge and Kegan Paul.

Grigg, David
1979 "Ester Boserup's theory of agrarian change: a critical review." Progress in Human Geography 3:64-84.
1980 Population Growth and Agrarian Change: An Historical Perspective. Cambridge, Mass.: Cambridge University Press.

Guillet, David
1981 "Land tenure, ecological zone, and agricultural regime in the Central Andes." American Ethnologist 8:139-158.

Hanks, L. M.
1972 Rice and Man: Agricultural Ecology in Southeast Asia. Chicago: Aldine.

Hallowell, A. I.
1943 "The nature and function of property as a social institution." Journal of Legal and Political Sociology 1:115-138.
1949 "The size of Algonkian hunting territories: a function of ecological adjustment." American Anthropologist 51:34-45.

Hardin, G.
1968 "The tragedy of the commons." Science 162: 1243-1248.

Harner, M. J.
1970 "Population pressure and the social evolution of agriculturalists." Southwestern Journal of Anthropology 26:67-86.

Herskovits, M. J.
1952 Economic Anthropology. New York: Norton.

Homans, G. C.
1960 English Villagers of the Thirteenth Century. Originally published in 1941. New York: Russell and Russell.

Izikowitz, K. G.
1951 "Lamet: hill peasants in French Indochina." Etnologiska Studier 17. Göteborg: Etnografiska Museet.

Johnson, Allen W.
1971 Sharecroppers of the Sertao. Stanford, Calif.: Stanford University Press.

Jones, E. L.
1974 "Environmental buffers on a marginal peasantry in Southern England." Peasant Studies Newsletter 4:13-16.

Land Tenure Center
1974 Agrarian Reform in Latin America: An Annotated Bibliography. Madison, Wis.: University of Wisconsin.

Leacock, Eleanor Burke
1954 "The Montagnais hunting territory and the fur trade." American Anthropological Association Memoir 78.
1963 "Introduction." Pp. i-xx in L. H. Morgan, Ancient Society. Cleveland, Ohio: World.

Leaf, Murray J.
1973 "Peasant motivation, ecology, and economy in Panjab." In K. Ishwaran, ed., Contributions to Asian Studies #3. Leiden: E. J. Brill.

Lee, Richard B.

1972 "!Kung spatial organization: an ecological and historical perspective." Human Ecology 1: 125-147.

1979 The !Kung San: Men, Women and Work in a Foraging Society. Cambridge, Mass.: Cambridge University Press.

Lee, Richard B., and Irven De Vore

1968 Man the Hunter. Chicago: Aldine.

LeRoy Ladurie, E.

1974 The Peasants of Languedoc. Urbana, Ill.: University of Illinois Press.

Lipton, M.

1968 "The theory of the optimizing peasant." Journal of Development Studies 4:327-351.

Lowie, Robert H.

1920 Primitive Society. New York: Liveright.

Marshall, Lorna

1976 The !Kung of Nyae Nyae. Cambridge, Mass.: Harvard University Press.

McArdle, Frank

1978 Altopascio: A Study in Tuscan Rural Society, 1587-1784. Cambridge, Mass.: Cambridge University Press.

McCay, Bonnie J.

1981a "Optimal foragers or political actors? Ecological analyses of a New Jersey fishery." American Ethnologist 8:356-382.

1981b "Development issues in fisheries as agrarian systems." Culture and Agriculture, Bulletin of the Anthropological Study Group on Agrarian Systems (11).

McCutcheon, Mary S.

1981 "Resource exploitation and the tenure of land and sea in Palau." Unpublished Ph.D. dissertation. University of Arizona.

Meggers, B. J.

1954 "Environmental limitation on the development of culture." American Anthropologists 56:801-824.

Meggitt, M. J.

1965 The Lineage System of the Mae-Enga of New Guinea. Edinburgh: Oliver and Boyd.

Moran, E. F.

1979 Human Adaptation: An Introduction to Ecological Anthropology. North Scituate, Mass.: Duxbury.

Morgan, Lewis Henry
1963 Ancient Society. Originally published 1877. Cleveland, Ohio: World.

Nell, E.
1972 "Boserup and the intensity of cultivation." Peasant Studies Newsletter 1:39-44.

Netting, Robert McC.
1965a "Trial model of cultural ecology." Anthropological Quarterly 38:81-96.
1965b "Household organization and intensive agriculture: the Kofyar case." Africa 35:422-429.
1968 Hill Farmers of Nigeria: Cultural Ecology of the Kofyar of the Jos Plateau. Seattle, Wash.: University of Washington Press.
1969 "Ecosystems in process: a comparative study of change in two West African societies." National Museum of Canada Bulletin (230).
1972 "Sacred power and centralization: aspects of political adaptation in Africa." In B. Spooner, ed., Population Growth: Anthropological Implications. Cambridge, Mass.: MIT Press.
1974 "Agrarian ecology." Annual Review of Anthropology 3:21-56.
1976 "What Alpine peasants have in common: observations on communal tenure in a Swiss village." Human Ecology 4:135-146.
1977 Cultural Ecology. Menlo Park, Calif.: Cummings.
1979 "Eine lange Ahnenreihe: Die Fortdauer von Patrilinien über mehr als drei Jahrhundert in einen schweizerischen Bergdorf." Schweizerische Zeitschrift für Geschichte 29:194-215.

Netting, Robert McC., and Walter S. Elias
1980 "Balancing on an alp: population stability and change in a Swiss peasant village." Pp. 69-108 in P. C. Reining and B. Lenkerd, eds., Village Viability in Contemporary Society. American Association for the Advancement of Science Selected Symposium 34. Washington, D.C.: American Association for the Advancement of Science.

North, D. C., and R. P. Thomas
1970 "An economic theory of the growth of the Western world." The Economic History Review 23:1-17.

Ottenberg, Simon
1958 "Ibo oracles and intergroup relations." Southwestern Journal of Anthropology 14:295-317.

Partsch, Gottfried
1955 Das Mitwirkungsrecht der Familiengemeinschaft in älteren Walliserrecht. Geneva.

Peterson, Nicholas
1975 "Hunter-gatherer territoriality: the perspective from Australia." American Anthropologist 77:53-68.
1979 "Territorial adaptations among desert hunter-gatherers: the !Kung and Australians compared." Pp. 111-130 in P. C. Burnham and R. F. Ellen, eds., Social and Ecological Systems. New York: Academic.

Powell, Sumner C.
1963 Puritan Village: The Formation of a New England Town. Middletown, Conn.: Wesleyan University Press.

Prosterman, Roy L.
1981 "El Salvador land reform is not a cruel hoax." Tucson Star, March 6.

Raftis, J. Ambrose
1964 Tenure and Mobility: Studies in the Social History of the Medieval English Village. Toronto: Pontifical Institute of Medieval Studies.

Rappaport, R. A.
1968 Pigs for the Ancestors: Ritual in the Ecology of a New Guinea People. New Haven, Conn.: Yale University Press.

Runge, C. F., and D. W. Bromley
1979 Property Rights and the First Economic Revolution: the Origins of Agriculture Reconsidered. Working Paper #13. Center for Resource Policy Studies. Madison, Wis.: University of Wisconsin.

Sahlins, Marshall D.
1961 "The segmentary lineage: an organization of predatory expansion." American Anthropologist 63:322-343.

Scott, James C.
1976 The Moral Economy of the Peasant: Rebellion and Subsistence in Southeast Asia. New Haven, Conn.: Yale University Press.

Silberbauer, G. B.
1972 "The G/wi Bushmen." Pp. 271-326 in M. G. Bicchieri, ed., Hunters and Gatherers Today. New York: Holt, Rinehart and Winston.

Skipp, Victor

1978 Crisis and Development: An Ecological Case Study of the Forest of Arden 1570-1674. Cambridge, England: Cambridge University Press.

Speck, F., and L. C. Eiseley

1939 "The significance of the hunting territory systems of the Algonkians in social theory." American Anthropologist 41:269-280.

Spooner, Brian, ed.

1972 Population Growth: Anthropological Implications. Cambridge, Mass.: MIT Press.

1974 "Irrigation and society: the Iranian plateau." In T. E. Downing and M. Gibson, eds., Irrigation's Impact on Society. Anthropological Papers of the University of Arizona, No. 25. Tucson, Ariz.: University of Arizona Press.

Spufford, Margaret

1979 Contrasting Communities: English Villagers in the Sixteenth and Seventeenth Centuries. London: Cambridge University Press.

Steward, Julian

1938 "Basin-plateau aboriginal sociopolitical groups." Bureau of American Ethnology Bulletin 120.

1955 Theory of Culture Change. Urbana, Ill.: University of Illinois Press.

1977 Evolution and Ecology. Urbana, Ill.: University of Illinois Press.

Turner, B. L., Robert Q. Hanham, and Anthony V. Portararo

1977 "Population pressure and agricultural intensity." Annals of the Association of American Geographers 67:384-396.

Udo, R. K.

1965 "Disintegration of nucleated settlement in East Nigeria." Geographical Review 55:53-67.

Van Gennep, Arnold

1960 The Rites of Passage. Chicago: University of Chicago Press.

Vayda, A. P.

1961 "Expansion and warfare among swidden agriculturalists." American Anthropologist 63:346-358.

Wallerstein, I. M.

1974 The Modern World-System: Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century. New York: Academic.

White, L. A.
1959 The Evolution of Culture. New York: McGraw-Hill.

White, Lynn
1962 Medieval Technology and Social Change. London: Oxford University Press.

Wiessner, D.
1977 "Hxaro: a regional system of reciprocity for reducing risk among the !Kung San." Unpublished Ph.D. dissertation. University of Michigan.

Wolf, Eric R.
1955 "Types of Latin American peasantry: a preliminary discussion." American Anthropologist 57:452-471.
1957 "Closed corporate peasant communities in Mesoamerica and Central Java." Southwestern Journal of Anthropology 13:1-18.

Yengoyan, Aram A.
1971 "The effects of cash cropping on Mandaya land tenure." In R. Crocombe, ed., Land Tenure in the Pacific. Melbourne: Oxford University Press.
1974 "Demographic and economic aspects of poverty in the rural Philippines." Comparative Studies in Society and History 16:58-72.

Cognitive Development in the First Years of Life

Katherine Nelson

Twentieth-century Americans concerned with child development--parents, educators, and scientists--have focused primarily on fostering intelligence through education and the measurement and improvement of cognitive achievements (Keniston, 1977). Over the years the specific issues have shifted from providing a universal secondary education and extending college education to all who could benefit from it to a concern over acquiring the basic skills of reading, writing, and arithmetic. In the past 15 years the realization that many children who entered school were not prepared to learn there led to the design and implementation of varying types of intervention programs beginning in the preschool years--Head Start, Follow-Through, Home Start, for example. The modest success of these programs in affecting intelligence scores (although their positive effects have been demonstrated in other ways, e.g., Schweinhart and Weikart, 1980) has led to the question of whether even earlier intervention might have more positive effects on development. In support of this presumption White (1975) has claimed that basic "competence" is established by age three and that therefore efforts at improving performance in later years can be ameliorative only. To be effective, optimal care must begin in infancy.

I am grateful for helpful criticism and comments on an earlier draft of this paper from William Kessen, Joseph Glick, Holly Ruff, and members of the Committee on Basic Research in the Behavioral and Social Sciences. I am indebted to Claire Kopp and Paul Harris for generously sharing their prepublication review chapters.

Meanwhile, another set of social forces has forced the issue of substitute care for infants and young children, as more and more mothers as well as fathers take jobs outside the home--almost half of the mothers of young children in the late 1970s. This movement has raised the urgent question of whether day care has negative effects on cognitive development for infants and young children. If, as White has claimed, children between six months and three years need an attentive one-to-one caretaker (preferably the mother) for optimal development, day care in the early years would seem to be detrimental. There are opposing views on this issue, however, that should not be rejected out of hand. Kagan et al. (1978), for example, report no negative effects--indeed no differences--on cognitive achievements between children placed in day care from early infancy and home-reared infants from comparable backgrounds.

As Kagan et al. point out, social policy decisions, such as whether to extend day care to a broad section of the population, are primarily matters of social values rather than logical choice. Yet, knowledge about the consequences of such decisions will surely contribute to wise decision making in the light of societal values. This paper reviews the kinds of knowledge that we have gained in recent years through basic research in early cognitive development. In so doing it reveals facts about development in the early years that have important implications for social policy toward children, although the research reviewed was not carried out with that end in mind.[1]

Before proceeding it is necessary to clarify basic terminology and the scope of this paper. Intelligence is a broad term generally taken, in our society to refer to the general level of intellectual competence displayed by an individual, a level that is presumed to remain stable over the life-span. It is indexed by the intelligence quotient (IQ), derived from a score on one of many different kinds of IQ tests--some administered to groups by paper and pencil, others given individually through oral questioning of a brief or extensive nature. There are many controversial issues related to the IQ concept, regarding its stability over the life-span, its malleability, and its relationship to the underlying concept of intelligence. The IQ is a psychometric concept, that is, it is derived from tests that measure differences between individuals. Although a child's IQ can be expected to remain relatively stable over time, compared with other children of the same age, the same child is also increasing in intelligence, that is, in adaptive competence, skills, and knowledge. The study of this kind of growth in intellectual competence, attributable to all children as they grow older (and to adults as well), is the study of cognitive development.

---

[1] I recognize that no social science research is value-free and that some of the researchers involved have indeed been motivated by the desire to shed light on, or to prove a point about, the effects of particular treatments. That fact does not, however, vitiate the main point of this paper.

Cognitive development has two facets: what the child knows at a given time and the processes by which she or he acquires more knowledge at a given time. These two facets define the different approaches of researchers in cognitive development: those who operate within an information-processing tradition--which is the dominant theoretical position of contemporary American cognitive psychology--and those who come from a structural, primarily Piagetian, tradition. The basic concern of all researchers in cognitive development is to identify changes in processes of intellectual functioning (such as visual attention and memory) and in the structure of knowledge.

Because, according to all theoretical accounts, cognitive functioning changes radically between one and three years of age--as the child becomes capable of using language, among other reasons--the study of cognitive development in infancy has been carried on quite independently of its study in later years (Bornstein and Kessen, 1979). For this reason, this review of research is confined primarily to developments in the first two years of life. Indeed, most of the research has been concerned with development in the first year and the age range between one and three years has been relatively little studied, except by those concerned with language development (Nelson, 1979).

In this paper I review three areas of early cognition that have seen major advances in recent years and show how our understanding of infancy and early childhood has changed thereby. These areas are the perception and cognition of the object world, social cognition, and early language development. I then briefly discuss their relationship to the identification and understanding of developmental delays and disorders and some of the possible implications for social policy. To set the stage for this review, I first place this research in a historical and theoretical context.

INFANCY: EARLY RESEARCH AND THEORY

The word infancy comes from the Latin root meaning "without language." The period of infancy is taken to be that period covering the first two years of life, before most children have gained productive control of and facility in using their first language. Piaget contributed the label sensorimotor stage to the first two years, emphasizing that, in contrast to later intellectual functioning that relies heavily on symbols, in this period of life the child negotiates the world through perception and action alone.

It is customary to note that Darwin was one of the first systematic observers of infant development, keeping a diary of his own son's progress (Darwin, 1877), thereby tracing the development of his perceptual skills and emotional expression. Toward the end of the 19th century diaries of infant development became common and were until very recently the primary source of data on early language development. During the 1920s and 1930s the major research emphasis in the area was on establishing norms of development through systematic and controlled testing, usually in institutes of child development such as those at the University of Minnesota, Harvard University, the University of California, Berkeley, and the Fels Institute in Ohio. The most ambitious and influential of these programs was no doubt that established at Yale University under Arnold Gesell, whose volumes on infancy and early childhood had enormous impact on psychologists as well as pediatricians and parents (Gesell, 1940, 1948; Gesell and Ilg, 1943). These volumes, like so much of the early work, were prescriptive as well as descriptive. They emphasized basic sensory development, the establishment of regularity, and the encouragement of independence. Little attention was given to language or to the development of intelligence.

Piagetian Theory

The growth of intelligence and the development of knowledge--that is, cognitive development--were the prime concerns of Jean Piaget (1896-1980), who, during the 1930s, was using the diaries he kept of his own three infants' first years to describe a presumed universal course in infant development and to develop his basic theory of genetic epistemology.[1] While this theory and the studies on which it was based have subsequently had an enormous impact on the study of early cognitive development, at that time they went largely unnoticed in the United States.

A brief summary of the main ideas of this theory as they apply to infant development is necessary to understanding the contemporary conception of infancy. Piaget viewed cognitive development as a slow process of the construction of reality on the part of each individual child (Piaget, 1954). In this view, the infant comes into the world with basic biological equipment, including a few reflexes such as those involved in sucking and looking. With these the child contacts and interacts with the world (at least a limited sphere of it) and gradually builds up sensorimotor action schemata through the interconnected processes of assimilation (of environmental objects to one's schemata) and accommodation (of the schemata to new objects). A sensorimotor schema is a behavioral organization that is applied in the presence of particular stimuli. For example, sucking begins as a reflex but becomes a schema as it assimilates not only breast but bottle, pacifier, thumb, blanket, and other objects. These come to be incorporated into the schema as "suckables." But in order to incorporate them, the schema must accommodate to their varying properties. It is from such beginnings, according to Piaget, that all cognition grows. Development is a process of forming more schemata, exercising and differentiating those one has, combining them into larger structures, and using them to establish general knowledge about the invariant properties of the world. Among the latter, Piaget places knowledge of objects at the center. During the course of the first two years the child is considered to be constructing general concepts about the nature of objects, space, time, and causality, all on the basis of the exercise of sensorimotor schemata.

---

[1] Genetic epistemology is a general term that Piaget uses to describe his theoretical work. He has been centrally concerned with the philosophy of knowledge, i.e., epistemology, and has chosen to study it by observing how children acquire knowledge structures (thus, genetic = development).

Between 18 and 24 months the child is said to emerge from the sensorimotor period and enter into the period of representational or preoperational thought. At this point the sensorimotor schemata become internalized and represented mentally. The symbolic function emerges and is evident in the child's acquisition of language, use of objects in symbolic play, and the capacity to image.

Piaget buttressed his account of these developments with detailed observations of his own three children's progress in infancy (much as Darwin had done). Spontaneous behaviors were supplemented by the construction of mini-experiments to test his own interpretations. For example, through a series of object-hiding games, in which small toys were hidden under cushions, rugs, or in the hand, Piaget tracked the development of what he termed the concept of the permanence of objects; that is, the idea that objects continue to exist when out of sight, an idea that the young infant apparently does not possess. Although based on only three subjects, Piaget's observations have proved to be highly reliable; the behaviors that he described have turned out to be typical of those displayed by most normal infants. Subsequently, it has proven possible to use these as the basis for developmental scales that have been applied to large samples of children (e.g., Uzgiris and Hunt, 1978).

However, the interpretation that Piaget gave to the data has not been universally accepted (e.g., see Kessen and Kuhlmann, 1970). Indeed, a great deal of the work on the perception and cognition of objects in infancy in the past 15 years has been generated by the desire to confirm or disconfirm the Piagetian theory of infant development.

Laboratory Methods

While Piaget was beginning to influence research in American universities, another source of interest in these areas came from the experimental laboratories there. In the latter part of the 1950s work undertaken by Fantz at Case Western Reserve University was revolutionary in its effect on subsequent research and the understanding of infant development (Fantz, 1958, 1961). In order to appreciate the impact of this work, it is necessary to reflect on the difficulty of obtaining information about the processing capacities of a nonverbal organism with limited responses--one that cannot be trained through rigorous methods involving food or water deprivation to produce a response on demand, as can lower animals in laboratory settings. Until the latter 1950s, investigators relied on physical measurements and direct observation of responses to stimuli. Studies of infant learning had been undertaken, such as the observation of learning to anticipate a feeding after a set interval, but the methods did not allow for fine estimations. The conclusions about the perceptual abilities of infants during this period were quite limited in comparison to what we now know. While Fantz's method seems extraordinarily simple in retrospect, it provided a breakthrough that led to much more sophisticated methodology and subsequent knowledge.

Fantz was interested initially in estimating the visual acuity of infants and for this purpose invented a procedure for determining visual preference. In this procedure two different stimulus designs (for example, a checkerboard and a plain gray square) are presented to the infant at a fixed distance while an observer looks through a peephole between the two stimuli at the infant's eyes and records the image that is reflected from the cornea, thus establishing a record of sequential visual fixations. The data gathered in this way showed the proportion of time that the infant fixated each stimulus and whether one was fixated significantly more frequently than another. If so, it could be claimed that the infant had a preference for that stimulus. More important, a preference indicated that the infant could discriminate between the two stimuli. By manipulating such factors as the width of stripes in a black and white grid compared with a gray pattern, Fantz was able to determine the degree of the infant's visual acuity and its development over the early months of life. This technique opened up a whole range of research problems in the area of visual perception, and subsequent refinements and variations on the method have produced a striking new picture of infants' basic capacities.

Soon after these developments, the use of film and later videotape was introduced to provide a permanent record of infants' responses for detailed analysis (e.g., Kessen, 1967). Videotaping techniques have truly revolutionized the area of infant cognition. Whereas natural observation of looking and acting remain the primary data base in this field, the ability to record these behaviors on tape and to review them in detail as often as necessary has infinitely improved our understanding of their significance. Studies based on these methods have both complemented, and in some instances conflicted with, the Piagetian-based description.

Attachment Theory

Another development during the 1950s that was strongly felt as an influence on basic research in the 1960s and 1970s was the theoretical contribution of John Bowlby (summarized in Bowlby 1969, 1973), known as attachment theory. This theory was based on both psychoanalytic thought and ethological concepts of sociobiological adaptation. It was also derived in part from empirical work demonstrating severe deleterious effects on infants who were separated from their mothers in impoverished institutional environments, a syndrome that came to be known as maternal deprivation. The theory suggested that the establishment of a mother-infant bond in the early months of life and its maintenance over the first two or three years would determine the quality of social adjustment in subsequent years.

There are two basic themes to attachment theory: separation and loss and secure attachment. Bowlby's general claim was that a series of basic biologically based events taking place in the mother-infant relationship were essential to establishing the mother-child bond characterized as attachment. Crying and clinging on the part of the infant with reciprocal soothing and holding by the mother constitute such basic interactions. When bonded, mother and child resist separation and seek each other on reunion. This notion both supplemented and to a large extent replaced the less complex and less attractive Freudian conception of a voracious id-driven infant bent on devouring the love object (the mother). Two of its controversial implications, which have inspired further research, are that it takes months to establish a secure attachment and that the mother (or at least the infant's primary caretaker) plays a vital and irreplaceable role in the child's early social and emotional development.

The other theme of attachment theory is that of separation and loss, in which the focus is on the child's ability to adapt to separation from the primary attachment figure. Here the data from hospitalized or institutionalized infants, who suffered extreme depression followed by rejection of the mother figure, were brought to bear. The theory implies that maternal separation or deprivation would be detrimental to development at least for the first two years. Much research has subsequently been carried out to substantiate or challenge these claims.

Linguistic Theory

Still another theoretical development with a major impact on research in infancy was the highly abstract linguistic theory of syntax known as transformational generative grammar, developed and promoted by Noam Chomsky (1957, 1965, 1976). Interest in language development had begun to emerge from its long period of neglect in the 1950s (e.g., Brown, 1958), but much of the impetus for the enormous creative surge in this area came from the implications of Chomsky's claims of a species-specific universal grammar. Its corollary, the innate language acquisition device (conceived to be a part of each child's native endowment), is supposed to take in the speech it hears, analyze it in terms of innate grammatical principles, and produce a correct grammar for the language being learned. This view stood in stark opposition to the prevalent view among psychologists of the 1950s that language, like all other behaviors, was learned through the general learning process of imitation and reinforcement.

Chomsky has had a major impact on philosophy and psychology as well as his own field of linguistics. His influence on developmental psychology has been indirect, in that it has been felt mostly in the negative reactions of researchers in the field to his claims. Not only is it obvious that an innate syntax-generating mechanism contradicts the basic assumptions of an environmentalist-oriented psychology (as American psychology has almost always been), but it is also in opposition to the constructivist approach of Piaget. Indeed, Piaget and Chomsky and their followers engaged in an extended debate a few years before Piaget's death, in which the issues of Chomsky's nativism and Piaget's constructionism formed the major points of contention (Piatelli-Palmarini, 1980).

Chomsky's theory was never grounded in developmental data and thus it has been open to developmentalists to test, to confirm, or disconfirm. Some central areas of controversy have been the implications that basic grammatical competence is innate and universal in the human species; that grammatical competence unfolds in a maturational sequence that does not depend on any particular environmental input (thus that there will be no effect on learning from the characteristics of the language heard); that language, or at least syntax, is independent of other cognitive or communicative developments. It is around these claims that much of the recent work on early language has been carried on.

Summary

These four developments, in the 1930s, 1940s, and 1950s--three theoretical (genetic epistemology, attachment theory, transformational grammar) and one methodological (the visual preference method)--set the stage for the exciting and far-reaching work of the 1960s and 1970s. Collectively this work has changed our view of the infant from a passive, unorganized, insensitive, unsocial, unaware, incompetent, unknowing organism to an active, sensitive, exploratory, social, and increasingly organized, knowledgeable, and competent organism. Although mother love and regular feeding were once felt to meet the needs of the infant, it is now known that babies thrive on intellectual and social stimulation of many kinds. The questions now are: What kinds? When? What in particular makes a difference?

THE OBJECT WORLD

Much of the research over the past 15 years has focused on the infant's relationship to objects--that is, the perception and cognition of objects. In this work objects are generally taken to be small and three-dimensional, although they may be two-dimensional representations, as in most of the research on visual perception. While the study of object perception and cognition may need no special rationale, its premiere place in theory derives from Piaget's description of sensorimotor development as well as from the exigencies of laboratory experiments based on concepts of information processing.

In Piaget's theory, action plays a central role. The child comes to know through action, and knowledge is represented primarily in terms of action schemes. In contrast, the work stemming from American laboratories was based on perception (the sensori side of the sensorimotor world). Most of the studies carried out in this tradition, although they may not have viewed the child as essentially passive, have treated the infant as passive in the experimental setting. This is partly an effect of methodology: It is difficult to study eye movements in a moving organism. As a result the methods that have been so successful with young infants are much less useful with infants of 10 months or more, who have gained some degree of mobility.

Perceptual Abilities

Until the 1950s there was considerable doubt as to whether the newborn could see or hear. Although this may seem to be an extreme statement, the response systems of the newborn are so irregular and unstable and the observational methods were so crude that no reliable data could be obtained to erase these doubts. At most there was an acceptance of the dictum that the baby exists in "one great booming buzzing confusion." This conception is not greatly different from Piaget's (1954) view of the infant in the first two months of life, who views the object world in terms of "pictures that can be recognized but that have no substantial permanence or spatial organization" (p. 2). These conceptions have been challenged by new knowledge. According to one prominent researcher, "More has been learned about infant perception in the last 15 to 20 years than in all previous years" (Cohen, 1979:894). The picture that has emerged of human development in the first six months is one of a highly active information-processing system that is undergoing rapid maturation, completing the biological development begun in the prenatal period, and is increasingly sensitive to environmental contingencies.

A brief indication of the focus of this work is the recent experimental work carried out in large part at laboratories at Yale University (Kessen), Harvard University (Kagan), the University of Minnesota (Salapatek), the University of Illinois (Cohen), the University of Denver (Haith and Campos), and Case Western Reserve University (Fantz and Fagan); this work has been devoted to tracing how the infant scans the object world, what infants can discriminate, and what they remember. The simple experimental paradigm of visual preference established a number of facts about infant perception. For example, visual acuity is quite poor at birth but increases dramatically over the first months of life. Young infants have presumably inborn preferences for patterned over unpatterned surfaces, for curved over straight lines, for faces versus nonfaces, and for certain configurations.

The later development of sophisticated infrared photographic techniques to study how newborn and older infants scan visual patterns (Salapatek and Kessen, 1966) showed that there is an apparent developmental shift over the first few months, from first scanning a figure globally, concentrating on the contour and angles, to a pattern of scanning the interior of the figure, concentrating on particular features. In the case of faces, whether schematic or real, the eyes tend to attract the most attention.

We know now that infants at birth perceive a great deal of the world around them and are actively engaged over the first months in organizing their perceptions. T. G. R. Bower (1974) has in fact attempted to demonstrate that the young infant has quite remarkable abilities to judge distance, to anticipate the trajectory of a moving object, and so on. Bower's claims (and those of others, such as Meltzoff and Moore, 1977, with respect to early imitation) challenge the Piagetian theory, since they claim that many advanced ideas about the world (such as the relationships of objects in space and to oneself) do not need to be constructed over time but are available to the organism from birth. These theoretical and factual disputes are still in contention and will no doubt provide the rationale for further research.

A variation of Fantz's procedure is the paradigm of familiarization, in which the infant is exposed over repeated trials to a particular stimulus and is then presented on a test trial with a pair of stimuli, the old one and a novel one. As with the simple preference design, differential looking at the novel stimulus is taken to indicate a discrimination between the two. This paradigm has proved useful in addressing questions in which an innate preference or ability was not the issue but memory or concept formation was.

Memory and conceptualization go beyond basic perceptual abilities, requiring the ability to organize information and hold onto it for later use. They are essential operations in intelligent action and are not specific to any particular perceptual modality, such as vision or hearing. Whatever the final outcome of the controversies over the nature of object knowledge in early infancy, all would agree that by six to eight months the infant has achieved the perceptual and motor skills that are basic to understanding the object world. The infant has achieved visual and auditory acuity, can integrate information from two modalities, can discriminate object characteristics (Ruff, 1980), and has control over increasingly fine motor movements necessary to the exploration of objects. At this point one can begin to investigate higher processes with some assurance that the information-processing system is reasonably mature in a physiological sense. Cognitive development in the period between 6 and 24 months is viewed differently by Piagetians and those in the information-processing tradition.

Infant Cognition: Information Processing and Piaget

The information-processing approach that dominates American cognitive psychology, and of which Fantz's early work might be said to be a precursor, emphasizes the representational aspects of cognitive functioning--that is, the sensori- or perceptually based side of sensorimotor intelligence. What interests researchers in this tradition are the development of visual perception, the formation of perceptual schemata for familiar patterns, the development of memory processes, and to some extent the development of motor skills. Jerome Kagan has been the most explicit theorist in this tradition. His most recent statements (Kagan et al., 1978), based on his conclusions from several longitudinal studies of infants in the first and second years, emphasize maturation in infant development and give a secondary role to experience. His basic theme is that the maturation of a process (for example, scanning the interior of a figure at 8 to 10 weeks or retrieving a previously stored schema at 8 to 10 months) makes possible the formation of a structure (the face schema or the relationship between events). Since development depends on the maturation of biological processes that are independent of environmental influence, Kagan in his recent work emphasizes the uniformity of basic developmental sequences and the resilience of the individual to overcome specific environmental deficiencies.

In contrast to Kagan's interpretation of the sequence of development, the Piagetian theme is epistemology, or coming to know. It is knowledge (not perception or memory) that is structured, and the medium of knowing in infancy is the coordination of action, not perception. Behavior is both a display of knowledge and a means of knowing. Piaget views development in infancy and throughout childhood in terms of movement through a series of stages. At each stage intelligence is organized within a system that operates according to certain principles in all aspects of its functioning. The operating principles differ from stage to stage. In contrast to Kagan's view, maturation alone is given little place in this system; growth is considered instead a continuous function of the adaptation between organism and environment. Adaptation involves the processes of assimilation and accommodation. The notion of a schema--a holistic and generalized framework--has come to be used quite widely in cognitive psychology. However, Piaget's schemata differ from Kagan's perceptual schemata: They are functions derived from interaction with objects, not simple observations of objects.

The research on cognition that has emerged from these two traditions overlaps in some respects but in others diverges. There are several areas in which some of the same phenomena and even the same terminology have been used to different ends. One of these is the object-hiding experiment.

Infants at about six months of age generally respond in a stereotypical way when an object is hidden. The object is usually a small attractive toy to which the infant's attention is easily drawn. The experimenter or tester ensures that the infant is watching, then covers the toy with a cloth or in some other way hides it. Instead of uncovering and retrieving the object, the infant, who has been watching intently, stops fixating the site where the object disappeared and turns to other things. This behavior is aptly characterized as "out of sight, out of mind." By about eight months, however, most infants will uncover the toy; Piaget views such behavior as indicative of the move to a new stage of cognition. But at this point another peculiar behavior pattern emerges. If there are two covers and two locations, A and B, and if the experimenter hides the toy first at A for a few trials and then switches to B, the infant, having watched the toy disappear at B will, instead of searching at B, go back to A.

The problem arises as to how to interpret and explain these highly reliable behaviors. A very large amount of research has now been carried out to test various competing hypotheses. Piaget's explanation is that the child must construct a concept of object permanence, and that prior to Stage IV in this sequence of development the infant does not search for the object because, since she or he can no longer see it, she or he believes that it has disappeared, that objects come and go in the world more or less randomly. A competing explanation, one that Kagan favors, is that the child's memory is insufficiently developed. (Other competing explanations, such as lack of motor skill to remove the cloth or inattentiveness, have been ruled out through experimentation. Development of a concept of object identity, put forth by Moore and his colleagues, or of spatial relations, favored by Bower, are quite close to the basic Piagetian concept and are not specially discussed in this illustrative review.)

Both Piaget's and Kagan's notions are deeply embedded in their theories of development; thus they are not simply alternative models of the object-hiding task but emerge from other systematic assumptions. Moreover, each has implications for basic views of the nature of intelligence. The interpretation of the object-hiding task, then, is a major confrontation between theories at a very early stage in development. Piaget's interpretation is organized around the developing structure of generalized knowledge about the logical relationships between objects, space, and time. Kagan's interpretation is in terms of the maturation of a particular process that is basic to intelligence--memory. Put this way, it is obvious that the two explanations do not compete but complement each other, one emphasizing structure and the other process. This is typical of the Piagetian versus non-Piagetian approaches to all questions of cognitive development.

In summary, the contribution of Piaget's work in this area has been very largely heuristic. He has contributed a vocabulary and a set of important observations that have spurred further work to unravel process explanations for the behaviors of normal infants and at the same time to establish a reliable sequence of their development. There is at present no final or generally accepted explanation for the developmental sequence of behaviors that is observed in the object-hiding paradigm. The phenomena continue to draw the interest of researchers from different schools of thought, in the conviction that behavior in this situation has broad implications for the development of intelligence and language.

Categorization in Infancy

One of the most dramatic changes in our conception of infancy and early childhood in recent years has involved the attribution to the infant of the ability to categorize. A long tradition of thought has asserted that young children were specifically deficient in the ability to form categories of objects based on rational principles of similarity. For example, demonstrations by both Piaget (Inhelder and Piaget, 1964) and Vygotsky (1962) have shown that, when presented with an array of objects--realistic or abstract geometric shapes--preschool-age children would use them to make functional groupings (such as putting a cow in a barn) rather than perceptual or categorical groupings (such as putting all the red objects together or all the animals). To theorists concerned with the development of logical principles of classification, the grouping of things on the basis of common attributes is felt to be more correct than grouping on the basis of functional associations. Thus the younger child was felt to be deficient in using principles of categorization. These deficiencies in behavior were thought to index deficiencies of thought.

It is important to note that this conclusion does not inevitably follow. For example, young children may be capable of forming groups of similar objects but fail to show the behavior because they did not understand what the experimenter or tester wanted. (Similar failures based on this kind of misunderstanding have been identified among illiterate adults in cross-cultural work; Cole and Scribner, 1974.) Or the young child might be capable of forming conceptual categories that are based on similarity but be misled by the concrete objects into interacting with them in a playful or imaginative way. The instructions that are usually given in this task are sufficiently vague as to encourage such behaviors.

The claim that young children (and of course infants) were conceptually deficient in this way made the development of language an overwhelmingly mysterious process, since language depends essentially on the ability to categorize perceptual input, as was emphasized by Roger Brown (1958). Sounds must be grouped into phonemic categories and then into words. Words must be matched with categories of objects, acts, properties, actors, and so on. To take a very elementary example, the term dog (or its baby form, doggie or bowwow) applies not to a single object but to a class of objects that share some (but not all) features in common. The child must be able to identify those animals that belong to the category dog and also those that do not. In other words, the child must be able to form an object category on the basis of similarity, precisely what young children were said not to be able to do in the standard object-sorting experiments. There have been many recent investigations of children's learning and use of terms such as doggie in the early stages of language acquisition, and the application of such terms on a categorical basis is not in doubt. Moreover, much more complex categories such as subject of a sentence or verb are formed and used in constructing sentences by two-year-olds. Still it could be argued, as those in the Chomskian school do, that language learning calls on specifically linguistic abilities that do not generalize to other behaviors.

A number of different lines of research, however, argue that this is not the case, that the categorization of perceptual input is a process that begins early in infancy and applies quite generally. One of the early demonstrations of this was carried out by Ricciutti (1965), who essentially extended the standard object-sorting experiment to the one-year-old level. He presented infants of 12, 18, and 24 months with a small group of abstract objects that differed along a single or multiple dimensions, then observed the children's behavior with them. He found a significant amount of both successive (acting on two similar objects in sequence) and simultaneous (putting two similar objects together in a group) comparison behaviors, although exhaustive group formation was not common even at two years. Nelson (1973) replicated this study using sets of realistic as well as abstract objects and showed that the form or function of the objects elicited grouping behaviors, while color did not. More recently Sugarman (1979) has adapted this general procedure to an intensive study of the development of grouping behaviors over the first three years. She has demonstrated that between one and two years children progress from acting on one kind of object when presented with two distinct sets of four objects each to comparing the two sets and making exhaustive sorts. That is to say, infants of 12 months do evidence categorically based behavior with regard to objects but do not spontaneously engage in complete logical classifications.

A different approach has been taken by Ross (1980) and by Cohen and Strauss (1979). Ross used the familiarization and preference comparison procedure developed in the infant perception research described earlier. She presented infants of 12, 18, and 24 months with several series of 10 different objects each, each series representing a different general category, such as men, animals, food, and furniture. After the presentation of the series the infant was presented with 2 new objects, 1 from the familiarized category and 1 from a novel category. In comparison with controls, even at 12 months the children preferred to look at the object from the novel category, thus demonstrating that they had represented the familiar category as a category and not as a series of discrete and unrelated stimuli.

Cohen and Strauss (1979) have extended this paradigm to younger infants (four to eight months) and have demonstrated the formation of a general "face" category even at seven months. Strauss demonstrated that such categories were based on the formation of a prototype in a way similar to that found with adults in other studies. Thus the perceptual categorization studies indicate very general cognitive abilities in quite young infants.

It no longer appears to be in doubt that infants are disposed to divide the perceptual world of objects into discrete categories even before they learn language and that these categories have the same basis as later categories do.

The disposition to treat stimuli categorically has also been demonstrated in the case of phoneme perception (Eimas, 1975; Morse, 1978); and in the area of color, infants have been shown to respond to colors in the same categorical way that adults do (Bornstein, 1976). They will, for example, treat a borderline blue (one that lies close to the spectral band classed as green) as "the same as" a focal blue rather than as green, to which it is physically closer. In both of these areas there appear to be built-in physiological bases for the nature of the discriminations made. Thus they differ from the apparently experientially based or culturally based categories that divide up the object world. Nonetheless, taken together these findings imply a view of the infant that is considerably more competent in sorting out the perceptual world than was thought to be the case only a few years ago.

Infant Memory

There have also been striking changes in our knowledge of the memory abilities of infants. Early studies of memory emphasized that babies would forget an episode or even a very familiar person (such as father) very rapidly. It was suggested that long-term memory did not develop until around three or four years of age. Recent work has challenged our assumptions in this respect, and although the total picture is now certainly far from clear, it presents us with a much more interesting and complex view of the infant mind.

The same type of familiarization and preference studies used in perception experiments has demonstrated both short-term and long-term memory abilities in very young infants. Indeed, the very fact that infants habituate attention to novel forms, colors, and phonemes indicates that they remember those forms, colors, or phonemes over a short period of time.

More important, it has become possible to investigate long-term memory by varying the delay interval between familiarization to a stimulus set and the preference test. Fagan (1976) familiarized five-month-old infants with a photograph of a man's face for two minutes and two weeks later found evidence that the infants retained memory for this face by presenting it with a novel photograph in a preference test. Other studies with five-month-olds (Cohen and Strauss, 1979) have shown that infants remember the shape, color, orientation, and size of an object on an immediate test. After 24 hours they still retain recognition of its shape. Retention of memory for faces appears to be greater than for other objects, a finding in line with the general preference for faces shown by infants.

These findings are somewhat puzzling in the light of the failure of infants of six or seven months to solve the object-hiding tasks described in a previous section. It seems probable that the coordination of skills remains an obstacle to infants' problem-solving capability. Another factor that may be involved is the type of memory that is required for different tasks. Familiarization and preference tests call on recognition memory, which involves only matching a present stimulus to some previously retained perception. The ability to recall an event in the absence of an adequate external cue is a different matter. It is difficult to obtain evidence of recall in a nonverbal organism. However, Ashmead and Perlmutter (1980) have obtained reports from parents of recall of events by children under 1 year. These reports rely on cues such as infants showing distress in situations reminiscent of a past painful one. Nelson and Ross (1980), using verbal memory reports, have shown that infants age 1 1/2 to 2 years may retain memories for specific events that occurred up to six months earlier and perhaps longer. In addition, in a controlled experiment DeLoache (1980) demonstrated that 2-year-olds could remember the location of a hidden object after a 24-hour delay. Taken together these recent reports indicate a far more reliable and powerful long-term recall memory in infants than was previously attributed to them.

Summary

What do these recent studies of cognitive processes add up to? Most researchers now believe that the human infant engages in quite complex information-processing and learning from a very early age. Although much neurological and motor maturation takes place over the first two years, basic capacities--of perceiving visual forms, colors, language sounds, objects, and spatial relations; of mentally categorizing them into groups of similar objects and events; and of remembering general concepts and specific events over long periods--all appear to be well developed in the last half of the first year. Some years ago a basic question was: Can infants learn? We now know that by two years of age children have established basic concepts and categories of thought and that their knowledge systems (or long-term memory) are well established. These achievements are of major significance to the further development of cognitive and linguistic functions.

THE CHILD'S CONCEPTION OF THE SOCIAL WORLD

Both the information-processing and the Piagetian approaches to understanding infant development largely neglect the social world, except as it is reflected in studies of face and mother perception. The social world is important in two ways to cognitive development: First, it is an object of knowledge itself; second, it provides support for the acquisition of knowledge in all realms--physical, social, and linguistic.

Just as there are phenomena that define the child's understanding of objects and object relationships, there is a well-established phenomenon around which theories of social cognition and development revolve. This is the syndrome consisting of fear of strangers and strange situations and protest at separation from the mother that appears at about eight months, peaks toward the end of the first year, and gradually declines thereafter. Attachment theory (Ainsworth et al., 1974; Bowlby, 1969) explains these developments in terms of the child's building up over the first nine months an attachment to the primary caretaker, usually the mother. In a threatening situation (such as the appearance of a stranger or in a novel situation) the infant clings to the safety of the mother and protests when she departs. Mother is thought to be the symbol of safety and comfort; because of the nature of the attachment bond, it is necessary for this primary bond to be established before the child can successfully extend its social and emotional bonds to others.

Modification of these claims is necessary, however. Recent research has shown that it is not necessary that a single attachment figure be established (Lamb, 1976), and theorists have disagreed with the emphasis on the exclusivity of the mother-child bond, which appears to be primarily a Western, middle-class conception. They have also questioned the importance of the mother-child bond as a model for later social relationships (Harre, 1974; Yandell, 1980). However, the focus of this paper is not on attachment theory per se but on the implications of these phenomena for the child's developing understanding of the social world--the development of social cognition. Therefore I consider how they have been interpreted in cognitive terms.

The findings reported in the previous section bear an important relation to the phenomena associated with attachment. Indeed, Kagan et al. (1978) and Bruner (1968), among others, have set forth a specifically cognitive interpretation of the development of stranger fear and separation protest. They believe that the infant begins to show fear of strangers at eight months because by that time she or he has built up mental representations or schemata of the primary caretakers and is able to distinguish them from unfamiliar people. The infant is also able to make distinctions between familiar and unfamiliar--and therefore possibly threatening--places. As the child begins to be able to imagine possibilities, she or he represents to herself or himself possible threats and therefore becomes distressed. However, with further development and experience she or he is able to allay these fears and to represent, on the basis of past experience, the probable return of the mother in her absence or the probable benignity of the strange place or face. This cognitive interpretation is still largely speculative on the basis of present knowledge. That the strange face, place, and separation phenomena may have a basis in cognitive processes rather than being solely an emotional development is buttressed by the fact that children in an Israeli kibbutz (Fox, 1977) or in infant day care (Kagan et al., 1978) exhibit similar behavior (including protest at separation from the mother) at similar ages, despite the fact that they are cared for by multiple caretakers who are not the mother.

Note that this interpretation applies concepts and findings from studies of physical or logical cognition to cognition of the social world. It has been common to assume that knowledge of the physical world (e.g., object permanence) is developed first and then extended to the social world. Some studies have shown, however (unsurprisingly), that most infants are able to demonstrate mother permanence, that is, to find the mother hiding behind a screen, before they demonstrate small object permanence (Bell, 1970). From this finding one may conclude that knowledge of the social world precedes and forms the basis for knowledge of the physical world, or at least that the two develop concurrently and independently (Gelman and Spelke, 1981).

But these attempts to explain social and object knowledge in terms of each other overlook the very real differences between the animate and inanimate as well as the distinctive contributions of the sociocultural environments, even apart from language and communication. Objects can be known through a fairly simple repertoire of behaviors: looking, handling, listening, and throwing, for example. In general, objects are under the control of people who use them or act on them in some way. They do not transform themselves independently or, if they do, their behavior is predictable: Glasses break if you drop them, bread becomes toast in the toaster, balls roll away, and so on. In contrast, the human world is far less stable and predictable. People move in unknowable directions and to unknown destinations. They react when acted upon in probabilistic, not deterministic, fashion. Even their movements are distinctive and are discriminable by infants from more mechanistic, less plastic movements, as Gibson et al. (1979) have recently shown.

Thus, to understand the social human world requires different predictive schemata than does the physical world (Glick, 1978). Obviously, the same cognitive processes must be involved--that is, perception, memory, and conceptualization. However, applying these processes to data from the observations of other people leads to quite different conclusions. For example, in many situations it is quite possible for the mother to disappear at one place and reappear at another--precisely what the infant must come to understand that objects cannot do if they are to achieve the concept of object permanence in Piaget's sense. Gelman and Spelke (1981) point out that even newborns make some distinctions between animate and inanimate objects, although their understanding of these distinctions apparently continues to develop throughout the preschool years.

In many ways, as these examples suggest, understanding the social world is a far more difficult and complex task than is understanding the world of objects. Yet infants manage to accomplish a great deal in this regard. As Richards notes (1974:85): "During the first year of postnatal development an infant is transformed from a social incompetent, totally dependent on the goodwill of adults, to a skilled social operator who is well able to hold his (or her) own in a wide variety of social situations."

This understanding and skill derive from participation in social activities with family members, either on a one-to-one affective partnership basis or as a participant in ceremonial acts such as greeting and leave-taking rituals (Harre, 1974). The latter, activities that are conventional, meaningful solutions to potential social problems--held to be the essence of the adult social world--have received little attention from students of early cognition, whereas the former have been the focus of numerous studies of early mother-child interaction (e.g., Schaffer, 1977; Stern, 1974; Trevarthan, 1977). This research has emphasized the degree to which mother and infant become tuned to each other during the first year, extablishing the form if not the content of a mother-child dialogue. Indeed, research by Conden and Sander (1974) has shown that even newborns synchronize their movements with the speech of adults. The role that mothers play in interpreting the sounds and expressions of their infants into meaningful exchanges has been the focus of research on infants of three to eight months. These exchange sequences are thought to form the basis for later language learning (see next section).

Although little is known about the child's organized knowledge of the social world in infancy, a number of studies have been aimed at discovering whether variations in maternal behavior affect the child's cognitive development. In summarizing the results of these studies, Ramey (1978:420-421) concluded:

> A relationship between maternal behaviors and infant competence or cognitive development is strongly suggested by the studies reviewed. Contradictions and controversy still exist, however. Many feel that standardized tests of infant development are inadequate tools for assessing the outcome of maternal influence upon cognitive development. . . . The difficulties inherent in observing, and thereby distorting and changing the very behaviors one wishes to study, have not been fully overcome, and there is a need to learn more about what these distortions may be. Little is yet known about how temperamental characteristics of infants combine with situational variables in affecting cognitive outcomes. The factor of direction of effects . . . is pertinent to this issue; are bright infants more stimulating and responsive to maternal behavior such that they serve as eliciting stimuli for greater interactive attempts by the mother, or do maternal behaviors serve to stimulate cognitive growth? The interaction of these two variables probably begins very early, making inferences about causation extremely difficult.
>
> Certain consistencies do seem clear, however. Maternal language seems to be a critical factor in infant cognitive development, both as to when and how it is used. Allowing the infant freedom to explore his environment seems to be important, as does sensitivity to the infant's needs and moods. The development of a positive social bond to another person seems to be related to cognitive development. Too much stimulation seems to be harmful. Clarke-Stewart (1973) has suggested possibilities for understanding the dynamic interplay between infant and mother: although the mother's stimulation of the infant seems to be the critical factor in the infant's intellectual development, the infant's stimulation of the mother may be most crucial in maintaining social contact.

As this summary indicates, there appears to be something that can be called optimal maternal care (Clarke-Stewart, 1973), involving verbal and social stimulation, positive effect, contingent responsiveness to the child, and play with the child that fosters optimal development in infancy. However, it should be noted that these factors are not the exclusive properties of mothers: They can be displayed by fathers, baby-sitters, or day care workers as well. Kagan et al. (1978) have concluded, on the basis of a large-scale longitudinal study of the effects of "quality day care" on all aspects of early development, that there are no differences in social or cognitive development between home-raised infants and those receiving day care. Thus the emphasis on the closeness and exclusivity of the mother-child bond that seems to be implied in the studies of mother-child interaction is probably misleading. Because of the structure of our society and the desire for controlled research designs, most of the relevant research has been concerned with mother-infant interaction (rather than father-infant or adult-infant), but we should be cautious about extending this practical result into prescriptive dogma. What these studies do indicate convincingly, however, is the kind of social support system that is optimal for cognitive development in the early years.

One of the areas toward which the results of the research on mother-infant interaction appears to be most relevant is that of the acquisition of linguistic and communicative skills, the focus of the following section.

LANGUAGE ACQUISITION

For many students of early development, advances in the understanding of language acquisition have been the most exciting of the past 15 years. Certainly the area has attracted some of the best minds in social science (e.g., George Miller, Roger Brown, Jerome Bruner), and it has also attracted many of the best graduate students. Conferences on language acquisition proliferate today, both nationally and internationally. Whereas 20 years ago there were scarcely half a dozen books on the subject, today there are hundreds and they continue to pour forth from the publishers.

What is the reason for all this activity? And why is research on language acquisition included in a review of infant development if infancy is indeed a period without language? The answer to the second question is that research has shown that the beginnings of language are rooted deep in the infancy period and that the process of development takes place over a very long period of time.

The answer to the first question is that language has been found to be a fulcrum for much of the development that takes place throughout the early years of life. The child's understanding of the object world and attachment to and understanding of social partners are intimately interwoven with the way that language is mastered. Thus the more that is discovered about the process, the more far-reaching the research to follow becomes.

Certainly one of the most important contributions of the "new" linguistics of the early 1960s is its emphasis on the structure of language and the central role of grammar in understanding and producing speech. Grammar is conceived as a set of rules that relate meaning to sound. A simple grammar describes how words can be combined to produce meaningful and grammatically "correct" sentences. Several efforts were made in the 1960s to write grammars that would describe the different stages that children go through in acquiring an adult-type grammar (see Brown, 1973, for a summary of these efforts). The typical research design at that time was one in which a group of investigators created and transcribed tape-recordings of two or three toddlers talking to their mothers in their homes at weekly or monthly intervals over a period of months and sometimes years. The transcriptions were then analyzed for the purpose of describing the steps by which children gain control over particular grammatical forms. An early observation was that children began to talk in two-word "sentences" at about 18 months and by four years had acquired most of the basic grammatical features of the language. Moreover, in many ways each child appeared to go through similar stages, that is, to acquire the same features in the same order.

These two characteristics--the speed of acquisition and the apparent universal order of development--appeared to support the claim put forth by Chomsky (1965) that in an important sense grammar is an innate capacity that unfolds according to a maturational scheme and that is not dependent on any specific kind of linguistic or cognitive input. In this view linguistic ability is independent of other kinds of cognitive ability. Lenneberg's book on the Biological Basis of Language (1967) appeared to support this position.

These claims did not go unchallenged, however. In response, researchers from psychology, linguistics, sociolinguistics, and anthropology undertook a variety of studies to test the implications of this position. In the process they examined the roots of language abilities in infant development along three main lines: cognitive development, communicative development, and the nature of the linguistic input, that is, the language the child hears. A subsidiary concern became the study of individual differences. I consider each of these lines briefly for the light they shed on cognitive as well as linguistic development.

The Cognitive Basis of Language

The most common formulation of the cognitive basis for language development is to assume that those developments in the first year of life that Piaget described in terms of sensorimotor intelligence form the foundation for the first words and sentences learned by the infant (see Brown, 1973). Several programs of research have investigated the validity of this assumption, attempting to relate the achievement of the concept of object permanence (described above), for example, to the emergence of syntax. These studies have been reviewed by Corrigan (1979), and in general the results have been mixed. Few studies have found a direct relationship between particular cognitive achievements and the onset of language. (Although there is some positive association between age of language achievement and performance on standard developmental tests, neither of these measures predicts well later intellectual performance; McCall, 1979.)

Bates and her colleagues (Bates, 1979; Bates et al., 1977) have attempted the most ambitious test of the relationship of cognitive abilities to language learning. Although they found no correlation between knowledge of object relationships and language measures, they did find fairly strong correlations between the ability to separate means and ends and measures of early language. They have proposed that the infant must achieve both the ability to subordinate the means to achieve a goal and the ability to imitate the production of others in order to acquire language. (Their theory does not imply a simple learning-by-imitation relationship, however.) According to their investigations, both of these abilities are generally achieved by the end of the first year and therefore may serve as a foundation for language development in the second year.

The most controversial questions in this area revolve around the issue of whether cognitive abilities in themselves are sufficient to enable a child to acquire language, or whether there are, as Chomsky claimed, specifically linguistic structures that must be called on. Can any intelligent creature learn a natural human language, or must one be humanly intelligent to do so? Since all children are human and therefore presumably have whatever species-specific equipment is required, this issue is difficult to resolve with empirical studies.

A different approach to these problems is to identify the processes that would have to be involved in language acquisition and to model the process via computer simulation to test whether the assumptions of the model are sufficient. Some efforts of this sort are under way but are still at a primitive stage. The acquisition of grammatical forms takes place between two and four years, and this period really lies outside the scope of this review. Further research along these lines can be expected to provide important clues to the cognitive foundations of language, however. For example, the work of Maratsos and his colleagues (Maratsos and Chalkley, 1981) is specifically designed to test whether nonlinguistic cognitive abilities are sufficient for the acquisition of grammar.

The beginnings of language clearly lie within the period surveyed. Children generally begin to understand words between 9 and 12 months of age, and they begin to produce a few recognizable words at around the first year. Toward the middle of the second year they begin to combine words into two-word sentences and to accelerate the acquisition of vocabulary. Analysis of the demands of these achievements in terms of the cognitive skills required reveals the importance of recognition memory; recall memory; the imitation of forms; and the categorization of sounds, objects, and relationships (Huttenlocher, 1974; Nelson, 1979). As discussed in the first section, these are all achievements of the first year of life. So far there have been no successful efforts to relate the development of these specific cognitive processes to the achievement of language. Future research would be expected to show that children who were more mature in these areas would be likely to begin to acquire and make progress in language learning earlier than children whose skills were less well developed and more uneven. The general skill approach to language acquisition is associated with the work of Bruner (1970) and his colleagues and Fischer (1980). However, so far no large-scale studies of these relationships have been forthcoming.

The Communicative Basis of Language

Much attention has been given to the fact that language does not unfold in a vacuum but must be prepared within the context of prelinguistic communicative exchanges. The studies of mother-infant interaction referred to in the last section are of relevance here. They have shown that (at least among the European and American middle class) mothers interact verbally with their young infants as though they were engaging in a conversation, even though the infants do not yet themselves speak. Thus the mothers establish a conversational structure involving the taking of turns that seems to be a precursor of the conversations the child is expected to engage in when she or he learns the appropriate words.

Another line of research shows that children begin to talk in highly structured and predictable communicative situations. Bruner (1975) speaks of the mother's setting up frames or formats within which the child gradually learns a part. It has been established that children usually begin to understand a few words related to simple names such as peekaboo or patty-cake (Benedict, 1976). Beyond this, it is apparent that the situations within which the child hears the language used determine what she or he learns and how she or he uses it (Nelson, 1981).

Enough is now known about the communicative context of learning to call into the question any model of language learning that assumes that the child is presented with a more or less random sample of speech in a neutral context. On the contrary, effective learning contexts are highly charged with effect and highly structured. Parents often quite literally put words into their children's mouths, not just at first but throughout the period of language learning (Berko Gleason, 1980; Schiefflin, 1979). While we need to know much more about the fine structure of the process of going beyond first words into sentences and complex grammatical forms, we know that these first steps are continuous with the cognitive achievements of the first year with respect to both the physical and social world. There is no reason to believe that they call on specifically linguistic processes. Whether later developments do depend on such processes is still an open question.

The Linguistic Input

Much attention has been paid by researchers interested in the acquisition of grammar to the form of the language that the child hears. Does the child need a particular kind of language model? Do different models provide better or worse learning environments? The issue is not one of whether parents use good or bad grammar but whether the language they use with their language-learning children is particularly effective as a teaching language. Many studies relevant to this issue have been carried out in the past 10 years, and they have uniformly shown that, when talking to 1- and 2-year-old children, adults use a style that is characterized by short sentences, simple vocabulary, many questions, and repetitions and is highly pitched and highly intonated in comparison with the speech style used with adults. It has been questioned as to whether this style can be appropriately characterized as simpler grammatically than speech ordinarily used with adults (which contains many more declarative sentences). Moreover, it is accepted that the perceived function of the style is to get and hold the attention of the child. (Similar characteristics appear in speech to pets, invalids, and foreigners.) Whatever the function of the style, it does appear to be conducive to learning by the child, as shown in recent studies that find positive correlations between many of these characteristics in mothers' speech in the second year and the children's language ability months later (Cross, 1978; Furrow et al., 1979; Wells, 1981).

Individual Differences

We have noted variations among children in cognitive processes and skills, variations in communication contexts, and variations among mothers (and presumably fathers) in language style. It is not surprising that children should vary in their language learning as well. What is of some interest is that there appear to be two common styles of learning that are roughly characterized as (1) _referential_ or _analytic_, in which the child learns many single words, especially object names, and constructs simple two-word sentences and (2) _expressive_ or _gestalt_, in which the child learns phrases, especially those that have pragmatic social uses, and puts together whole clumps of preformed sentences instead of or in addition to the building up of single-word combinations. The former style is associated with picture-book reading and object orientation; the latter with social contexts. While it seems probable that the communicative context of learning may be the effective mechanism determining which style the child uses, it is also possible that differential maturation of brain structures may be involved. In either case these distinct approaches to the tasks have implications both for understanding the process as a whole and for helping those who have difficulty in acquiring language.

Implications

Learning language is a long and difficult task, calling on the child's basic cognitive abilities and building on his or her knowledge of both the physical world and the world of social relationships. It is not discontinuous with cognitive development in the first year but combines all that has been achieved and carries the child beyond prelinguistic status. The language user has acquired a cognitive and communicative tool of great power, but it would not be possible to do so if he or she had not already developed considerable cognitive and communicative skills. Moreover, it would not be possible if the environment did not provide suitable models and contexts for learning. It is an interactive system with the potential for development in a variety of directions. Language carries the child beyond infancy, but developments in infancy have made its acquisition possible.

DEVELOPMENTAL DELAYS AND DISORDERS

In recent years much effort has gone into the study of children whose cognitive development is slow or disordered in some way. This research relates to basic research on normal cognitive development in two ways: First, knowledge of normal development provides a basis for assessing deviant development; and second, the study of deviant cases can shed light on the factors that are important in normal development.

With respect to the first aspect, there are several important contributions, as a recent review by Kopp (in press) highlights. Most evaluations of infants and young children have relied on the uses of psychometric tests of mental development. However, research such as that reviewed above makes possible the use of more specific psychological tests and measures that relate delays and disorders to specific areas of functioning. For example, Fagan's (1978) tests of infant memory have been used with infants with Down's syndrome (previously termed mongolism and usually involves severe retardation) as well as with normal infants. Moreover, he has shown that performance on infant memory tasks in the first year predicts intellectual performance at five to six years. In other research, attention and visual scanning have been useful techniques for assessment. The use of Piagetian-type assessments of sensorimotor intelligence with blind infants (Fraiberg, 1974) and infants with Down's syndrome has also helped to identify particular areas of functioning that are affected. As research advances, it is possible to relate developmental disorders to theoretically important causal mechanisms, enabling us at least to understand the disorder and perhaps to deal with it more effectively.

Aside from the use of more sensitive and theoretically defensible tests, the understanding of infant development that has emerged over the past 15-20 years has led also to a revised view of research and treatments of infants "at risk," that is, those whose biological development or environmental conditions suggest the possible emergence of later disabilities. In particular, as Kopp points out, the model of development as neither the independent unfolding of a preformed program nor the cumulative effect of environmental inputs but rather as a continuously changing interaction has been widely accepted. Inherent in this model accepted by both the information-processing school and the Piagetians is the notion of an active organism that seeks to adapt within the environment with whatever resources are at hand. Thus an infant with some type of physical or mental disability can be expected to follow a different pattern of development from the normal one, using his or her available capacities in a fundamentally adaptive way. That is to say, the interactive model views handicaps not in terms of deficiencies, but in terms of differences that are brought about by the nature of the interactive process.

An example of a specific contribution of basic research to understanding disorders of development is discussed by Kopp in the case of the acquisition of language by children with Down's syndrome. These infants apparently develop slowly but normally in terms of their sensorimotor skills. They face problems, however. in developing language and abstract symbolic functioning in general. Some earlier studies showed that parents of children with Down's syndrome used less complex language when talking to them than parents of normal children used with children of the same age; it was hypothesized on this basis that the retarded children were being deprived of an adequate language model. However, the psycholinguistic research had shown (as discussed earlier) that adults generally adapt their speech to young children who are learning language in the direction of shorter sentences, simpler vocabulary, and so on. When the speech of parents talking to their Down's syndrome children was compared with that of parents talking to normal children in the same stage of language development (instead of at the same age), it was found that they were very similar. What had appeared to be a parentally produced handicap for the retarded child turned out to be simply a realistic adaptation on the part of parents to the actual level of the child's functioning. Thus no valid conclusion could be drawn that Down's syndrome children were suffering deprivation in the language area.

Another area in which knowledge of normal development has been useful is studies of preterm infants and blind infants, both of whom often have difficulties with mother-child attachment behavior. Studies of attachment behavior in normal children have shed light on the source of these difficulties, for example, lack of early contact with preterm infants and lack of eye contact with blind infants. Knowledge of these effects has led to some successful efforts to ameliorate the difficulties by encouraging interaction with preterm infants and educating mothers to recognize the sources of their own distress.

Discerning the usefulness of research on developmental disorders for understanding normal development is not at first so self-evident. Three examples drawn from Kopp's review may suffice to make the point. She cites Gouin-Decarie's (1969) research on the development of infants who were deformed as the result of ingestion of thalidomide by their mothers during pregnancy. The question of interest in this research was the degree of dependence of the concept of object permanence on the child's ability to act on objects. Thalidomide children who lacked arms could not interact with objects in the normal way but did develop normal object concepts, thus suggesting that the development of this basic cognitive structure was not as dependent on action per se as had been thought.

Another case in point is the development of Down's syndrome children. These children suffer varying degrees of mental retardation beginning in the first year of life, falling progressively behind their peers, apparently not advancing much beyond the 2- to 3-year-old level of mental age. Yet despite this retarded developmental course, these children develop the same cognitive structures that normal infants do, according to Kopp. The pace of development is slower, but the form of behavior is unaffected. This conclusion has implications for normal development in that it reinforces the interpretation that sensorimotor behaviors "have a firm biological basis that reflects strong evolutionary pressures, and are distorted only in the wake of profound organismic damage" (Kopp, in press).

This is not to say that Down's syndrome infants show no differences from normal children. They appear to have difficulty in discriminating or processing subtle or complex signals such as those displayed in communicative situations. Further studies of these aspects of development should shed more light on the kinds of signals children normally rely on in processing and communicating information.

A third area with implications for normal development is research on the developmental course of preterm infants. Infants born prematurely have been the focus of a large number of studies in recent years as it has become recognized that outcomes for infants with very low birth weight are often very poor. The deleterious effects of treatment or the lack of it in the perinatal period are often but not always long-lasting. Many conditions affect the degree to which an infant suffers lasting consequences. One of the prominent conditions is the economic status of the family. Preterm births are much more frequent among low-income families, and long-term outcomes for children in poverty areas are much less positive. Certainly nutritional and other health factors play a large role in these relationships. Further study of specific cases should reveal much more about what is needed for optimal development in the face of handicaps associated with preterm birth. This in turn is likely to shed more light on the generally wide disparity in intellectual functioning between children from homes in poverty and those from more advantaged backgrounds.

IMPLICATIONS FOR SOCIAL POLICY ISSUES

In the light of the advances in knowledge about early development in recent years, a number of policy issues can be addressed much more intelligently than was the case 20 years ago. Three of these will be considered here: infant day care, intervention programs, and parent education. Each involves questions of what conditions are optimal for children's development and how best to achieve them.

The focus of this paper has been on cognitive development broadly conceived to include development of cognitive processes, such as memory, conceptualization, and problem-solving skills (what we usually think of as intelligence) and knowledge of the world, physical and social and symbolic. These two aspects are integrally bound together; greater facility in exercising intelligence generally leads to increased knowledge, while greater knowledge makes possible a more powerful application of intelligence. What are the optimal conditions under which this system develops?

It should first be stressed that cognition, even so broadly defined, is only one aspect of the development of the young child, one that cannot in reality be separated from the child's health, affective development, and personal-social attitudes and ties. We have seen that affective and social development have been found to be of considerable significance to the infant in establishing a secure base from which to explore the world. The question of health has not, however, been considered specifically thus far. The fact that the health of the infant is no longer of prime concern in our society is itself worthy of note. As Newson and Newson (1974:55) put it: "The ability to view infant care practices in any other light than whether they help the infant in his basic struggle to live is . . . something peculiar to our own century, and within this century, to the technologically advanced countries in which infant and child mortality rates have now been reduced to comparatively insignificant proportions."

It seems obvious that the first requirement for public policy is to see that adequate health care is extended to all infants, prenatally as well as postnatally. Recent research has revealed the lasting effects on infants of poor nutrition, alcoholism, and drug addiction in pregnant women (e.g., Streissguth et al., 1980). Mental deficiency in the progeny is one of the common outcomes of these conditions. Severe malnutrition in infancy, unless corrected early, also leads to mental retardation. Since much of our knowledge of these effects has been obtained very recently, it seems likely that there are other, more subtle effects of specific deficiencies or insults that have not yet been identified. Clearly the first concern of those charged with the care of infants must be the maintenance of adequate nutrition and health care.

We also know, from the many studies of children in institutions who receive adequate food and medical attention but "fail to thrive," that adequate physical care is not sufficient to meet the needs of infants, emotionally, cognitively, or physically (e.g., Dennis, 1938; Spitz, 1950). These studies have raised the issue of whether infant day care is an appropriate approach to the problem of meeting the needs of working mothers for substitute child care. They have raised the spectres of institutionalism and maternal deprivation. As pointed out earlier, the studies stemming from attachment theory appeared to support the interpretation that infants need the close and constant care of a mother figure early in life, thus making day care for very young children an undesirable alternative to home care, but more recent research has modified these conclusions.

There are two sides to this problem: One is consideration of what realistic alternatives for infant care exist; the other involves the lessons to be drawn from research on infancy. Between a third and a half of the mothers of children younger than age three in this country are now employed (Hoffman, 1979); a significant percentage of the rest are on welfare. For both of these groups the choices for child care are highly restricted. If infant day care is not available to working mothers, the alternative, as has often been shown, is usually an ad hoc baby-sitting arrangement, sometimes no care at all, or care by a barely older sibling, as young as age five or six. Moreover, mothers on welfare are frequently unable to meet even the minimum nutritional and health needs of their infants, much less their emotional and cognitive needs. Therefore, in a real sense for a vast population the question is not whether infant day care is optimal but whether it is less disadvantageous than the alternatives. Let us consider the light that basic research can shed on this issue.

Kagan's large day care project (recently summarized in Kagan et al., 1978) compared low-income children from Chinese or Caucasian families in day care continuously from about 4 to 29 months with home-reared controls from comparable backgrounds on a large number of variables at different ages, including attachment and social, cognitive, and language development. They concluded (p. 261):

> The complete corpus of data does not offer much support for the view that quality day care outside the home has an important effect on the young child's development. . . . It is surprising that 3,500 hours of regular contact with other young children had little influence on degree of apprehension, responsiveness, or the disposition to be aggressive or cooperative with an unfamiliar child. . . . Attachment to the mother and rate of cognitive development, the two critical concerns of American parents, did not appear to be altered by the day care experience. The children in group care treated the mother as if she were the primary agent of nurturance and no mother indicated . . . that her child had become either estranged or indifferent to her psychological pressures for socialization. The assessments of language, memory, and perceptual analysis failed to reveal any obvious advantages or disadvantages to the day care experience. . . . Our answer to the central question that provoked the investigation is that a child's attendance at a day care center staffed by conscientious and nurturant adults during the first two and one-half years of life does not seem to produce a psychological profile very much different from the one created by rearing totally in the home.

This conclusion is not uncontroversial and the authors themselves enter several cautions with regard to it. First, it applies to high-quality day care with a high ratio of nurturant caretakers to infants and other quality inputs (for full description see Kagan et al., 1978). Second, they emphasize that day care is not risk-free: Children usually contract more illnesses in group care than at home; shy or withdrawn children may not thrive in a group environment; and, unless given special attention, language development may lag. However, on the whole this controlled study found none of the differences in cognitive and affective development that might be expected.

Kagan et al. used the most sophisticated measures of cognitive development that they could glean from previous research in making their comparisons. Their findings of no differences attributable to group versus home environments on these measures were a surprise to them and have raised controversy among researchers at large. But while their findings are important, they do not by any means indicate that no environmental conditions have an effect on early cognitive development. Certainly children seem more immune to fairly large environmental variations than the most enthusiastic environmentalists would have believed, and some reformulations of beliefs about early development may be in order.

Two of the major theories that have inspired much of the basic research in early cognitive development would predict that specific minor variations in environments would have little effect: Chomsky's theory of innate language structure maturing according to a specific timetable and Piaget's theory of organism-environment interaction leading to a universal construction of basic logical structures. Ethologically based attachment theory would predict a disruption of mother-child bonding under group arrangements, but the finding that this did not happen is explained by the assumption that this biologically based system is more protected from disruption than previously thought.

What then are we to make of the various research outcomes traced above, some of which have shown variation in the development of these systems as a function of environmental input? As noted earlier, recent research on social development, of which the work of Kagan is one example, has demonstrated that infants do not suffer from having more than one major caregiver or attachment figure. Grandparents, fathers, baby-sitters, or day care workers who provide a secure and nurturant base for the child are clearly not a disadvantage and may actually be an advantage for development.

With respect to cognitive and linguistic development, it is clear that there is variation among children with respect to both the rate and type of development in these areas. It is probably the case that much of this variation is a function of innate factors: differences in basic abilities, in tempo, in disposition, and so on. These differences clearly exist. It is also the case that, particularly in the area of language development, these differences interact with and are affected by the environment provided by caregivers. Kagan emphasizes, as did Clarke-Stewart (1973) and Nelson (1973), that an optimal environment is one that allows children to explore and initiate and that reacts to their attempts to express themselves rather than artificially restricting, confining, or ignoring them. Children with different dispositions (or of temperament or intellect) approach situations in different ways. Thus what any child gets from a given environment is a function not only of the environment but also of what he or she brings to it. Thus the conclusion is not that there is not significant variation among children that is affected by environmental conditions, but rather that a major difference between two fairly rich environments has no predictable effect on all children. We also know that children from less adequate environments often emerge from infancy seriously deficient in basic cognitive and linguistic functions.

This conclusion is in line with both theory and research that rely on an interactive model, whether Piaget's model or more process-oriented models of cognitive development. The outlines of the general position can be summarized as follows: Children come into the world with a genetic program that contains both universal species-specific dispositions (to learn language, develop memory capacities, categorize perceptual input) and individual characteristics (of processing abilities, temperament, physical and motor skills). In addition, infants at birth are already the product of the prenatal organism-environment interaction and may have been affected deleteriously by that experience (Kopp, in press). During the first two years of life those dispositions unfold according to a maturational timetable that dictates the course of development of the central nervous system and motor skills, but it is modulated by those individual characteristics that determine how the infant meets the environment and what he or she finds on those encounters. The general program appears to be largely protected from environmental variation, and the individual characteristics tend to ensure that a wide variety of environments are accepted in similar ways. The nature of the interaction ensures both variation among children in similar environments and similarity among children in different environments. It also implies intraindividual variation over time in measures of cognitive development, language, and so on, because different individual interactions derive different inputs from the same environment at different points in development.

These considerations also point to the source of the great resilience found among many children who have suffered severe deprivation or malnutrition or abuse in early childhood (Rutter, 1978). These children can apparently overcome the effects of these problems when the situation is corrected because they can call on modes of interaction that lead to a rapid catch-up or compensation. However, some children are not so fortunate, for whom the effects of deleterious conditions are long-lasting. It is also the case that some environments have negative effects on all children exposed to them.

Given these considerations intervention programs of the Head Start type for preschoolers may be viewed in a different light. The need for such programs, whether family-oriented or school-oriented, is an indication that social policy with respect to infants has failed in a basic way by not ensuring an environment for all children in which they can develop optimally during the early years of life. Head Start and other intervention programs are predicated on the concept of catch-up and resilience, of making up deficiencies. As noted above, while there are many spectacular and well-documented cases of resilience in the face of major insults, not all children recover from negative experiences. As long as some children emerge from infancy with poorly established cognitive structures and linguistic skills, the need for catch-up programs will persist. And if followed through into the early school years, the programs will be moderately effective (Schweinhart and Weickert, 1980; Zigler and Trickett, 1978) for many children. But would it not be better to put some of that effort into ensuring for all children the minimal conditions needed to support the development of basic cognitive abilities, skills, and knowledge in the first years of life?

Part of the establishment of such conditions surely would involve an effective program of parent education. The lessons from basic research on cognitive development in infancy are easy to teach and easy for prospective parents to apply. A great deal of money is not needed to implement these lessons in the home. Good nutrition, accessibility to objects of varying kinds (not necessarily expensive toys), an interactive social partner or partners, responsiveness to the child's initiatives, "framing" encounters in repetitive routines that the child can understand--these are some of the simple things that parents can be taught to provide. These examples do not exhaust the list of things that could be usefully taught by any means, of course, but my purpose here is not to provide a curriculum guide.

One of the issues that has loomed large in efforts to change parental behavior through the popular press is that of making children smarter. Surely research on cognitive development must have something to say on this issue. Is this not also the aim of intervention programs?

This issue is fraught with misinterpretation and misunderstandings. IQ tests are the standard means by which smartness is assessed in our society. As discussed above, IQ tests, when applied to stable populations in standard environments, show considerable stability over time, especially during the school years. That is, measures of intelligence relative to one's peers tend to be stable within an individual during childhood and into adulthood. This is not to say that there are no intraindividual variations; many children go up or down 15 or more IQ points for no readily identifiable reason. On the whole, though, one can predict future IQ scores fairly well from about age six. Up to age six tests predict less well, probably because they measure different abilities (Honzik, 1976; McCall, 1976).

Recent studies of adoption have shed light on the extent to which different environmental factors directly affect IQ scores. The answer seems to be in general not very much (Willerman, 1979). Children resemble their biological parents in IQ and do not resemble their adoptive parents. The correlation coefficient of child and adoptive parent IQ hovers around zero. This issue is complicated, however, by the fact that it has been demonstrated, in studies of black adopted children in white families, of adopted Korean orphans, and of a massive intervention program in Milwaukee (Heber, 1978), that improved environments can have substantial positive effects on IQ. That is, children can be made "smarter" than they otherwise would be expected to be if left in their previous environments. Note that all these studies are based on populations at risk for deleterious effects on development. One can surmise that the intervention provided a more optimal environment within which the child's innate capacity could develop to an optimal degree. It did not make the child smarter but enabled her or him to become as smart as she or he was designed to be.

Is there a limit to these effects? Can an average child in a good environment be made brilliant by appropriate intervention? There is no good evidence that this can be done. Clearly, children can be force-fed knowledge. They can be taught to read at two; some can be taught to read Latin and Greek at three, as John Stuart Mill was. But the history of such prodigious efforts to produce prodigies is not a happy one. The goal of making children smarter is probably not a realistic one to hold out to parents, and it may even be harmful. As noted in the first three sections of this paper, there is much to be learned and mastered by the infant and young child that is not strictly cognitive, although it calls on cognitive processes--such things as how to participate in social interaction and social ceremonies and the uses of language, fantasy, and play. The period of infancy and early childhood is one in which the child, through exploration, guided and unguided, applies his or her own maturing abilities to come to know the world. The fact that the infant-child does not display the cognitive skills that older children do--of reading, writing, classifying, thinking logically--does not mean that she or he is not exercising her or his own cognitive skills in important ways.

Educating parents to provide an environment that will encourage and nourish these developments should be an important policy goal. Encouraging parents or policy makers to believe that we can raise children's overall intellectual level through such efforts is probably unrealistic. It does not seem unrealistic, however, to believe that such efforts might be reasonably effective in ameliorating particular negative practices and conditions, especially if they were combined with appropriate services such as health care and, where needed, quality day care.

CONCLUSION

It is not easy to summarize all that has been achieved in advances in our knowledge of early cognitive development over the past 20 years. The following is an attempt to bring out the most important points.

From birth, infants are actively engaged in taking in and organizing information about the environment. This is an interactive process in which what is already known guides what can be learned next. There is no simple process of stamping in knowledge from outside or simple maturation but rather a complex interaction between a maturing and increasingly knowledgeable organism and an environment that offers different opportunities for learning at different points in development. Parents and other caretakers who understand the dynamic nature of the developmental process and who are sensitive to changing needs and competencies, encouraging exploration and inquiry but offering also routines and frameworks that can be mastered and understood, provide optimal environments for early development.

Basic cognitive developments in infancy are under the control of strong biological forces and are not easily derailed by specific environmental conditions. Individuals use whatever resources are available to achieve basic competence and often show considerable resilience in overcoming temporary deprivation or handicaps. There are conditions, however, both biological (e.g., deafness or blindness) and environmental, that prevent the achievement of normal and basic cognitive structures and skills. At this point we do not know the details of which conditions are necessary for normal development, although we know a great deal about common interactive processes. We do know that some conditions that might be thought to be detrimental (group day care, for example) are not; and we also know that some children from disadvantaged homes who are apparently physically intact have fallen behind intellectually by the end of the infancy period and have become candidates for preschool intervention programs. That at least some of this retardation is unnecessary is apparent from the results of such massive intervention programs as the Heber (1978) study. Moreover, while programs such as Head Start have had considerable effectiveness in many areas, the results of these programs do not hold out hope that deficiencies of the early years can easily be compensated for in later development.

Basic research has contributed enormously to our understanding of these propositions. There is every reason to expect that further research will help us to understand the environmental side of the interactive process in more detail and in so doing indicate what every child needs for optimal intellectual functioning.

REFERENCES

Ainsworth, M. D., S. M. Bell, and D. J. Stayton
1974 "Infant-mother attachment and social development: 'socialization' as a product of reciprocal responsiveness to signals." In M. P. M. Richards, ed., The Integration of a Child into a Social World. London: Cambridge University Press.

Ashmead, D. H., and M. Perlmutter
1980 "Infant memory in everyday life." In M.

Perlmutter, ed., Children's Memory. San Francisco: Jossey-Bass.

Bates, E.
1979 The Emergence of Symbols: Communication and Cognition in Infancy. New York: Academic Press.

Bates, E., L. Benigni, I. Bretherton, L. Camaioni, and V. Volterra, V.
1977 "From gesture to first word: on cognitive and social prerequisites." In M. Lewis and L. Rosenblum, eds., Interaction, Conversation, and the Development of Language. New York: Wiley.

Bell, S.
1970 "The development of the concept of object as related to infant-mother attachment." Child Development 41:291-311.

Benedict, H.
1976 "Language comprehension in 10-to-16-month-old infants." Doctoral dissertation. Yale University.

Berko Gleason, J., and S. Weintraub
1980 "Input language and the acquisition of communicative competence." In K. E. Nelson, ed., Children's Language. Volume 1. New York: Gardner Press.

Bornstein, M. H.
1976 "Infants are trichomats." Journal of Experimental Child Psychology 21:425-445.

Bornstein, M. H., and W. Kessen, eds.
1979 Psychological Development from Infancy: Image to Intention. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Bower, T. G. R.
1974 Development in Infancy. San Francisco: Freeman.

Bowlby, J.
1969 Attachment. New York: Basic Books.
1973 Separation. New York: Basic Books.

Brown, R. W.
1958 Words and Things. New York: The Free Press of Glencoe.
1973 A First Language: The Early Stages. Cambridge, Mass.: Harvard University Press.

Bruner, J. S.
1968 Processes of Cognitive Growth: Infancy. Worcester, Mass.: Clark University Press.
1970 "The growth and structure of skill." In K. Connolly, ed., Mechanisms of Motor Skill Development. London: Academic.

1975 "The ontogenesis of speech acts." Journal of Child Language 2:1-20.

Chomsky, N.

1957 Syntactic Structures. The Hague: Mouton.

1965 Aspects of the Theory of Syntax. Cambridge, Mass.: MIT Press.

1976 Reflections on Language. Glasgow: Fontana.

Clarke-Stewart, K. A.

1973 "Interactions between mothers and their young children: characteristics and consequences." Monographs of the Society for Research in Child Development 38:6-7, 153.

Cohen, L. B.

1979 "Our developing knowledge of infant perception and cognition." American Psychologist 34:894-899.

Cohen, L. B., and M. S. Strauss

1979 "Concept acquisition in the human infant." Child Development 50:419-424.

Cole, M., and S. Scribner

1974 Culture and Thought. New York: Wiley.

Conden, W. S., and L. W. Sander

1974 "Neonate movement is synchronized with adult speech: interactional participation and language acquisition." Science 183:99-101.

Corrigan, R.

1979 "Cognitive correlates of language: differential criteria yield differential results." Child Development 50:617-631.

Cross, T. G.

1978 "Mothers' speech and its association with rate of linguistic development in young children." In N. Waterson and C. Snow, eds., The Development of Communication. New York: Wiley.

Darwin, C.

1877 "A biographical sketch of an infant." Mind 11: 286-294.

DeLoache, J. S.

1980 "Naturalistic studies of memory for object location in very young children." In M. Perlmutter, ed., Children's Memory. San Francisco: Jossey-Bass.

Dennis, W.

1938 "Infant development under conditions of restricted practice and of minimum social stimulation: a preliminary report." Journal of Genetic Psychology 53:149-158.

Eimas, P.

1975 "Speech perception in early infancy." In L. Cohen and P. Salapatek, eds., Infant Perception: From Sensation to Cognition. Volume 2. New York: Academic Press.

Fagan, J. F.

1976 "Infants' delayed recognition memory and forgetting." Journal of Experimental Child Psychology 21:425-445.

1978 "Infant recognition memory and early cognitive ability: empirical, theoretical, and remedial considerations." In F. D. Minifie and L. L. Lloyd, eds., Communicative and Cognitive Abilities: Early Behavioral Assessment. Baltimore: University Park Press.

Fantz, R.

1958 "Pattern vision in young infants." Psychological Record 8:43-49.

1961 "The origin of form perception." Scientific American 204:66-72.

Fischer, K. W.

1980 "A theory of cognitive development: control and construction of a hierarchy of skills." Psychological Review 87:477-531.

Fox, N.

1977 "Attachment of kibbutz infants to mothers and metapelit." Child Development 48:1228-1239.

Fraiberg, S.

1974 "Blind infants and their mothers: an examination of the sign system." In M. Lewis and L. Rosenblum, eds., The Effect of the Infant on its Caregiver. New York: Wiley.

Furrow, D., K. Nelson, and H. Benedict

1979 "Mothers' speech to children and syntactic development: some simple relationships." Journal of Child Language 6:423-442.

Gelman, R., and E. Spelke

1981 "The development of thoughts about animates and inanimates: implications for research on social cognition." In J. H. Flavell and L. Ross, eds., The Development of Social Cognition in Children. New York: Cambridge University Press.

Gesell, A.

1940 The First Five Years of Life. New York: Harper & Bros.

1948 Studies in Child Development. New York: Harper & Bros.

Gesell, A., and F. L. Ilg

1943 Infant and Child in the Culture of Today. New York: Harper & Bros.

Gibson, E. J., C. J. Owsley, and J. Johnson

1979 "Perception of invariants by 5-month-old infants: differentiation of two types of motion." Developmental Psychology.

Glick, J.

1978 "Cognition and social cognition: an introduction." In J. Glick and K. A. Clarke-Stewart, eds., The Development of Social Understanding. New York: Gardner Press.

Gouin-Decarie, T.

1969 "A study of the mental and emotional development of the thalidomide child." In B. M. Foss, ed., Determinants of Infant Behavior. Volume 4. London: Methuen.

Harre, R.

1974 "The conditions for a social psychology of childhood." In M. P. M. Richards, ed., The Integration of a Child into a Social World. London: Cambridge University Press.

Heber, F. R.

1978 "Sociocultural mental retardation: a longitudinal study." In D. Forgays, ed., Primary Prevention of Psychopathology. Volume 2: Environmental Influences. Hanover, N.H.: University Press of New England.

Hoffman, L. W.

1979 "Maternal employment: 1979." American Psychologist, 34:859-865.

Honzik, M. P.

1976 "Value and limitations of infant tests: an overview." In M. Lewis, ed., Origins of Intelligence. New York: Plenum Press.

Huttenlocher, J.

1974 "The origins of language comprehension." In R. L. Solso, ed., Theories in Cognitive Psychology: The Loyola Symposium. Potomac, Md.: Erlbaum.

Inhelder, B., and J. Piaget

1964 The Early Growth of Logic in the Child. New York: Harper & Row.

James, W.

1890 Principles of Psychology. New York: Henry Holt.

Kagan, J., R. B. Kearsley, and P. R. Zelazo
1978 Infancy: Its Place in Human Development. Cambridge, Mass.: Harvard University Press.

Keniston, K.
1977 All Our Children: The American Family under Pressure. New York: Harcourt Brace Jovanovich.

Kessen, W.
1967 "Sucking and looking: two organized congenital patterns of behavior in the human newborn." In H. W. Stevenson, E. H. Hess, and H. L. Rheingold, eds., Early Behavior: Comparative and Developmental Approaches. New York: Wiley.

Kessen, W., and C. Kuhlmann
1970 Thought in the Young Child. (Originally published 1962.) Chicago: Chicago University Press.

Kopp, C.B.
in press "Risk factors in development." In M. Haith and J. Campos, eds., Infancy and the Biology of Development. Volume II of P. Mussen, ed., Manual of Child Psychology. New York: Wiley.

Lamb, M.
1976 The Role of the Father in Child Development. New York: Wiley.

Lenneberg, E. H.
1967 Biological Basis of Language. New York: Wiley.

McCall, R. B.
1976 "Toward an epigenetic conception of mental development in the first three years of life." In M. Lewis, ed., Origins of Intelligence. New York: Plenum Press.
1979 "The development of intellectual functioning in infancy and the prediction of later I.Q." In J. D. Osofsky, ed., Handbook of Infant Development. New York: Wiley.

Maratsos, M. P., and M. A. Chalkley
1981 "The internal language of children's syntax: the ontogenesis and representation of syntactic categories." In K. E. Nelson, ed., Children's Language, Volume II. New York: Gardner Press.

Meltzoff, A. N., and M. K. Moore
1977 "Imitation of facial and manual gestures by human neonates." Science 198(Oct. 7):75-78.

Morse, P. A.
1978 "Infant speech perception: origins, processes, and alpha centauri." In F. D. Minifie and L. L. Lloyd, eds., Communicative and Cognitive

Abilities--Early Behavioral Assessment. Baltimore: University Park Press.

Nelson, K.
1973 "Some evidence for the cognitive primacy of categorization and its functional basis." Merrill-Palmer Quarterly 19:21-39.
1979 "The role of language in infant development." In M. H. Bornstein and W. Kessen, eds., Psychological Development from Infancy: Image to Intention. Hillsdale, N.J.: Lawrence Erlbaum Associates.
1981 "Individual differences in language development: implications for development and language." Developmental Psychology 17(2): 170-187.

Nelson, K., and G. Ross
1980 "The generalities and specifics of long-term memory in infants and young children." In M. Perlmutter, ed., Children's Memory. San Francisco: Jossey-Bass.

Newson, J., and E. Newson
1974 "Cultural aspects of childrearing in the English speaking world." In M. P. M. Richards, ed., The Integration of a Child into a Social World. London: Cambridge University Press.

Piaget, J.
1954 The Construction of Reality in the Child. New York: Basic Books.

Piatelli-Palmarini, M., ed.
1980 Language and Learning: The Debate between Jean Piaget and Noam Chomsky. Cambridge, Mass.: Harvard University Press.

Ramey, C. T., D. C. Farran, F. A. Campbell, and N. W. Finkelstein
1978 "Observations of mother-infant interactions: implications for development." In F. D. Minifie and L. L. Lloyd, eds., Communicative and Cognitive Abilities--Early Behavioral Assessment. Baltimore: University Park Press.

Ricciutti, H.
1965 "Object grouping and selective ordering behavior in infants 12 to 24 months old." Merrill-Palmer Quarterly 11:129-148.

Richards, M. P. M.
1974 The Integration of a Child into a Social World. London: Cambridge University Press.

Ross, G.
1980 "Categorization in 1 to 2-year-olds." Developmental Psychology.

Ruff, H. A.
1980 "The development of perception and recognition of objects. Child Development 51:981-992.

Rutter, M.
1978 "Early sources of security and competence." In J. S. Bruner and A. Garton, eds., Human Growth and Development: Wolfson College Lectures. Oxford: The Clarendon Press.

Salapatek, P., and W. Kessen
1966 "Visual scanning of triangles by the human newborn." Journal of Experimental Child Psychology 3:155-167.

Schaffer, H. R.
1977 Studies in Mother-Infant Interaction. London: Academic Press.

Schiefflin, B. B.
1979 "Getting it together: an ethnographic approach to the study of the development of communicative competence." In E. Ochs and B. B. Schiefflin, eds., Developmental Pragmatics. New York: Academic Press.

Schweinhart, L. J., and D. P. Weikart
1980 Young Children Grow Up: The Effects of the Perry Preschool Program on Youths through Age 15. Ypsilanti, Mich.: The High/Scope Press.

Spitz, R. A.
1950 "Anxiety in infancy: a study of its manifestations in the first year of life." International Journal of Psychoanalysis 31:138-143.

Stern, D. N.
1974 "Mother and infant at play: the dyadic interaction involving facial, vocal and gaze behavior." In M. Lewis and L. A. Rosenblum, eds., The Effect of the Infant on its Caregiver. New York: Wiley.

Streissguth, A., F. Landesman-Dwyu, J. C. Martin, and D. W. Smith
1980 "Teratogenic effects of alcohol in human and laboratory animals." Science 209:353-361.

Sugarman, S.
1979 "Scheme, order and outcome: the development of classification in children's early block play." Ph.D. dissertation. University of California at Berkeley.

Trevarthan, C.
1977 "Descriptive analyses of infant communicative behavior." In H. R. Schaffer, ed., Studies in Mother-Infant Interaction. London: Academic Press.

Uzgiris, I. C., and J. McV. Hunt
1978 Assessment in Infancy: Ordinal Scales of Psychological Development. Champaign, Ill.: University of Illinois Press.

Vygotsky, L. S.
1962 Thought and Language. Cambridge, Mass.: MIT Press.

Wells, G.
1981 "Apprenticeship in meaning." In K. E. Nelson, ed., Children's Language, Volume 2. New York: Gardner Press.

White, B.
1975 The First Three Years of Life. Englewood Cliffs, N.J.: Prentice-Hall.

Willerman, L.
1979 "Effects of families on intellectual development." American Psychologist 34:923-929.

Yandell, D. L.
1980 "Sociability with peer and mother during the first year." Developmental Psychology 16:355-361.

Zigler, E., and P. Trickett
1979 "IQ, social competence and evaluation of early childhood intervention programs." American Psychologist 33:789-798.

From Experimental Research to Clinical Practice: Behavior Therapy as a Case Study

G. Terence Wilson

INTRODUCTION AND OVERVIEW

A Historical Note

Behavior therapy is a relatively new psychological approach to the assessment and treatment of emotional and behavioral disorders and educational problems. A defining characteristic of behavior therapy, which has served to set it apart from other psychological therapies, is the deliberate, continuing attempt to base treatment concepts and methods on psychology as an experimental science. Although some of the procedures that are part of current behavior therapy had been described in the first half of this century, it was not until the late 1950s and early 1960s that behavior therapy emerged as an explicitly formulated, systematic approach to assessment and treatment. At this early stage of development behavior therapy was derived directly from what was referred to as modern learning theory--more specifically, the principles and procedures of classical and operant conditioning.

Developed by Pavlov at the turn of the century, classical conditioning refers to the procedure of associating a neutral or conditioned stimulus (i.e., one that does not automatically evoke a reflex or response) with an unconditioned stimulus (i.e., one that naturally elicits a response). Pairing a conditioned stimulus with an unconditioned stimulus eventually results in the ability of the conditioned stimulus alone to elicit the response. In classical conditioning, events or stimuli that precede behavior control the response. For example, Mowrer and Mowrer (1938) conceptualized the relatively common childhood problem of enuresis (bed-wetting) as a failure of certain stimuli (bladder cues) to elicit a response (waking) so that the child can rise and urinate appropriately. Accordingly, they devised a treatment based on classical conditioning principles. They constructed an apparatus that consists of a urine-sensitive pad that is connected to a bell. When the child urinates, the bell automatically rings, serving as an unconditioned stimulus for waking. Bladder distension that precedes the unconditioned stimulus eventually elicits waking prior to urination and the sounding of the alarm. The procedure results in control of urination and permits the child to sleep through the night without urinating. Subsequent research has shown that this bell-and-pad technique is a safe, effective, and inexpensive method for eliminating enuresis (Ross, 1981).

With the rise of behaviorism in the United States early in this century and its emphasis on the objective study of psychology, learning processes and procedures such as classical conditioning became a dominant focus of research. Among the notable figures who expanded our knowledge of the laws of learning and laid the foundations for the growth of behavior therapy in the 1950s were Hull, Spence, Tolman, Guthrie, Mowrer, and Miller. The experimental studies of Mowrer and Miller were particularly significant, since they investigated fear responses and avoidance behavior. Their findings directly influenced the development of techniques for treating anxiety-related disorders, which are discussed below. Another prominent behaviorist, Skinner, contributed the detailed analysis of operant conditioning, in which a major portion of behavior can be shown to be a function of its consequences to the organism.

Research on conditioning and learning principles, conducted largely in the animal laboratory, had become a dominant theme in experimental psychology in this country during the years that followed World War II (Kimble, 1961). In the traditions of Pavlov and Skinner, basic researchers in this area were committed to the scientific analysis of behavior using the laboratory rat and pigeon as their prototypic subjects. Applications of learning principles to clinical disorders, however, including Mowrer and Mowrer's extension of conditioning principles to the treatment of enuresis, were isolated and sporadic efforts that had scant impact on psychotherapy.

These early extensions of learning principles to human problems received no formal attention in part because conditioning principles, which had been demonstrated with animals, were rejected as too simplistic and irrelevant to the treatment of complex human problems. Under the influence of traditional psychodynamic therapies, conditioning treatments were rejected as superficial, mechanistic, and naive. A schism developed between academic experimental and clinical psychologists. The former were trained in scientific methods, with an emphasis on controlled experimentation and quantitative measurement. The latter concerned themselves with projective tests, intrapsychic inferences, and speculative hypotheses about the unconscious motives behind human behavior. Students enrolled in doctoral clinical programs gained their clinical training in a medically oriented setting such as a state hospital, a Veterans Administration center, or a psychiatric clinic. Essentially abandoning their academic training in psychology as an experimental science, students in these psychiatric settings learned to apply quasi-disease analogies in treating mental illness. Behavior therapy was developed as a formal attempt to bridge this gap between laboratory and clinic.

Modern Origins of Behavior Therapy

Foremost among the modern origins of behavior therapy were the following developments. In 1958, Wolpe, a South African psychiatrist, published the classic text, Psychotherapy by Reciprocal Inhibition, in which he introduced several novel treatment methods based on the principles of Pavlovian conditioning, Hull's learning theory, and his own research on the elimination of experimentally produced "neurotic" reactions in laboratory animals. In England, Eysenck (1959) defined behavior therapy as the application of modern learning theory to psychiatric disorders. Behavior therapy was said to be a scientific approach based on experimentally demonstrated methods that were more effective than traditional psychotherapy. The latter was characterized as unscientific in nature, based on speculative theories and procedures, and lacking any acceptable evidence of its efficacy. In sum, according to Eysenck, behavior therapy was an applied science, the defining feature of which was that it is testable and falsifiable. A testable theory is one that can be specified with sufficient precision so that it can be subjected to experimental investigation. A theory that is falsifiable is one that specifies experimental conditions that could result in the theory being disproved or falsified. Eysenck argued that in contrast to learning theory, psychoanalysis was too vaguely formulated to be really testable and that it was impossible to identify conditions under which it could be falsified.

In the United States, Skinner (1953:373) reconceptualized psychotherapy in behavioristic terms:

> The field of psychotherapy is rich in explanatory fictions. Behavior itself has not been accepted as subject matter in its own right, but only as an indication of something wrong somewhere else. The task of therapy is said to be to remedy an inner illness of which the behavioral manifestations are merely "symptoms" . . . the condition to be corrected is called "neurotic," and the thing to be attacked by psychotherapy is then identified as a "neurosis." The term no longer carries its original implication of a derangement of the nervous system, but it is nevertheless an unfortunate example of an explanatory fiction. It has encouraged the therapist to avoid specifying the behavior to be corrected or showing why it is disadvantageous or dangerous. By suggesting a single cause for multiple disorders it has implied a uniformity which is not to be found in the data. Above all, it has encouraged the belief that pyschotherapy consists of removing inner causes of mental illness, as the surgeon removes an inflamed appendix or cancerous growth or as indigestible food is purged from the body. . . . It is not an inner cause of behavior but the behavior itself which--in the medical analogy of catharsis--must be "got out of the system."

Operant conditioning principles and procedures were rapidly extended by Skinner's students to the modification of diverse clinical and educational problems.

In sum, although these different pioneers in behavior therapy did not share an identity of views, they were drawn together in their conviction that effective psychological therapy must be based on the content and method of experimental psychology. (For a detailed history of behavior modification, see Kazdin, 1978a.)

Current Conceptions

Toward the end of the 1960s the theoretical and research bases of behavior therapy began to be expanded by drawing more broadly on areas of experimental research and psychological theory beyond classical and operant-conditioning principles. Increasingly, behavior therapists turned to current developments in areas such as social, personality, and developmental psychology for different ways of conceptualizing their activities and as a source of innovative therapeutic strategies.

Particularly noteworthy in this regard was Bandura's social learning theory, with its emphases on vicarious learning (modeling), symbolic processes, and self-regulatory mechanisms (Bandura, 1969). The earliest behavioral formulations of Wolpe, Eysenck, Skinner, and others had eschewed cognitive concepts in their analyses of behavior change, choosing to focus more narrowly on overt, observable responses. This appeal to conditioning as opposed to cognition stemmed from the early emphasis on principles from the animal conditioning laboratory and the initial reaction of behaviorism and behavior therapy against the mentalistic concepts of traditional psychodynamic approaches. Bandura's theory and research had a major influence on the practice of behavior therapy in clinical and educational settings. Examples of the use of modeling and self-control strategies in promoting psychological change are discussed later in this paper. The important point to note here is that these influential contributions of Bandura and other leading figures who tackled the theoretically challenging phenomena of unobservable or cognitive determinants of behavior and self-regulatory processes (e.g., Kanfer and Phillips, 1970) derived from basic laboratory research using human subjects.

Throughout the 1970s behavior therapists incorporated cognitive processes and procedures into their therapeutic armamentarium. In many instances this recent "cognitive connection" in behavior therapy has been a far cry from the original behavioristic formulations of Wolpe, Skinner, and other pioneers. New treatment strategies were generated, alternative theoretical interpretations of existing methods were proposed, and novel analyses of specific clinical disorders (e.g., cognitive models of depression) were introduced. Once again basic psychological research furnished the mainsprings of many of these conceptual and clinical developments. Among other sources of psychological research, attribution theory (a prominent part of contemporary cognitive-social psychology) and information-processing concepts had significant impacts on the nature of behavior therapy (Mahoney, 1974; Rosenthal, 1982). In 1974 Dember remarked that "psychology has gone cognitive."

Social, developmental, and experimental psychology all showed the profound influence of cognitive conceptualizations. That behavior therapy followed this cognitive trend, given its self-conscious link to psychology as an experimental science, is not surprising. By the late 1970s the term cognitive behavior modification had become a fixture in the field (Meichenhaum, 1977).

In sum, there is no monolithic approach to behavior therapy. Contemporary behavior therapy is marked by a diversity of views, a broad range of heterogeneous procedures with different theoretical rationales, and open debate about the adequacy of existing conceptual foundations. Despite these differences among behavior therapists with respect to specific theoretical frameworks, a common core of fundamental assumptions characterize the field. The two basic characteristics are a psychological model of human behavior, which differs fundamentally from the traditional intrapsychic, psychodynamic, or quasi-disease model of mental illness, and a commitment (in principle if not always in practice) to an applied science approach, to maintaining a link to the methods and substance of psychology as an experimental science. This approach has the following characteristics: (1) an explicit, testable conceptual framework; (2) treatment that is either derived from or at least consistent with the content and method of experimental clinical psychology; (3) therapeutic techniques that can be described with sufficient precision to be measured objectively and be replicated; (4) the experimental evaluation of treatment methods and concepts; and (5) the emphasis on innovative research strategies that allow rigorous evaluation of specific methods applied to particular problems instead of global assessment of ill-defined procedures applied to heterogeneous problems.

Of course, there is not, nor can there be, a simple isomorphic relationship between psychological research and clinical practice. Behavioral practitioners have necessarily borrowed concepts and treatment methods from other therapeutic approaches; moreover, following their counterparts from alternative theoretical orientations, they have developed their own clinical lore, much of which owes nothing to experimental research. Lacking sufficient information and guidelines from scientific research, behavior therapists frequently are obliged to adopt an informed trial and error approach to difficult or unusual problems. It must also be emphasized that beyond scientific knowledge and prowess, the behavior therapist must possess personal and professional qualities similar to those demanded by other psychotherapeutic approaches. So-called nonspecific factors such as warmth, empathy, caring, and sound clinical judgment are necessary but not sufficient ingredients of most therapeutic change (Wilson and Evans, 1977). The distinguishing feature of behavior therapy is the degree to which the findings of basic psychological research have had an indelible influence on clinical thinking and practice in behavior therapy--on the conceptual models of abnormal behavior and its modification, on specific assessment and intervention strategies, and on methodological advances in the evaluation of treatment outcome. The following sections of this paper provide illustrative examples of this crucial link between experimental research and effective therapeutic applications.

SELECTED EXAMPLES OF THE RELATIONSHIP BETWEEN PSYCHOLOGICAL RESEARCH AND THERAPEUTIC PRACTICE

Behavior therapy is broadly applicable to a wide range of problems. In addition to clinical psychology and psychiatry, areas of successful application include education, rehabilitation, and even medicine (Kazdin and Wilson, 1978). Indeed, it can be argued that behavioral treatment methods are applicable to a far wider range of clinical disorders than traditional psychodynamically oriented psychotherapies (Rachman and Wilson, 1980). The following sections present summary analyses of the role of basic research in behavioral treatment of selected problems. These problems have been chosen to convey the breadth of behavioral applications across widely differing disorders in different populations (including adults and children) as well as the personal and societal significance of the problems on which psychological research has had a salutary effect.

The paper concludes with the presentation of a comprehensive model of the overall relationships among different levels of psychological research and therapeutic applications. The model indicates the necessity of different types of research for an adequate applied science of clinical treatment to be realized.

Anxiety Disorders:
Fears, Phobias, Obsessions, and Complusions

Specific fear reduction techniques that are commonly used to treat the full range of neurotic anxiety disorders were derived directly from conditioning research on animals. These techniques include systematic desensitization, flooding, in vivo exposure, and response prevention, and collectively they provide striking evidence of the value and potential of basic research for therapeutic applications.

Fears and Phobic Disorders

Among the most effective and widely used methods for overcoming phobic disorders are systematic desensitization, flooding, and in vivo exposure.

Systematic Desensitization This technique was developed in its present form by Wolpe (1958) on the basis of his research on the experimental induction and elimination of "neurotic" fears in cats. He showed that conditioned emotional responses could be eliminated by feeding animals gradually closer to the locus of the original fear conditioning. Accordingly, Wolpe formulated his reciprocal inhibition principle, which stated that anxiety-eliciting stimuli could be permanently neutralized if "a response antagonistic to anxiety can be made to occur in the presence of the anxiety-evoking stimuli so that it is accompanied by the complete or partial suppression of the anxiety response" (p. 71). In applying these laboratory findings to the development of a practical clinical treatment technique, Wolpe adapted from Jacobson (1938) a technique called progressive relaxation training as a means of producing a response that is incompatible with anxiety. Briefly, this consists of training people to concentrate on systematically relaxing the different muscle groups of the body, which results in lowered physiological arousal and a comfortable feeling of calmness.

In another extrapolation from basic laboratory research findings to the therapeutic situation, Wolpe found that imaginal representation of stimulus conditions seemed to be as effective in eliciting anxiety as their actual occurrence. This development was of major importance in aiding the therapist to deal with anxiety reactions that could not be easily controlled or directly dealt with in the therapist's office. The patient is asked to imagine anxiety-producing scenes in a carefully graded fashion. Hierarchies of anxiety-eliciting situations are constructed, ranging from mildly stressful to very threatening items, which patients are instructed to imagine while they are deeply relaxed. Each scene is repeated or the hierarchy adjusted, until the person can visualize the scene without experiencing anxiety. Only then does the therapist present the next item of the hierarchy. When appropriate, the patient is usually instructed to engage in graduated performances of previously feared activities in the real world, a procedural variant known as in vivo desensitization.

Systematic desensitization has been the most extensively researched technique among the various psychological therapies (e.g., Rachman and Wilson, 1980). A reassuring feature of the massive clinical and research literature on systematic desensitization is that, while there is lively debate over the theoretical mechanisms that are responsible for anxiety reduction, its therapeutic efficacy has been generally accepted. A strongly positive evaluation of systematic desensitization has remained consistent over the years. In a major review of the evidence on systematic desensitization in 1969, Paul was able to conclude that "the findings were overwhelming positive, and for the first time in the history of psychological treatments, a specific treatment package reliably produced measurable benefits for clients across a broad range of distressing problems in which anxiety was of fundamental importance. 'Relapse' and 'symptom substitution' were notably lacking, although the majority of authors were attuned to these problems" (p. 159). Leitenberg (1976) later commented that "it seems safe to conclude that systematic desensitization is demonstrably more effective than both no treatment and every psychotherapy variant with which it has so far been compared" (p. 131).

Systematic desensitization has proved effective for the full range of phobic reactions, ranging from simple, circumscribed complaints (e.g., fear of small animals, heights, and so on) to complex and debilitating disorders, such as agoraphobia. A unique feature of this evidence is that it comes from studies that employed sophisticated controls, unprecedented in psychotherapy research, to exclude alternative explanations of the observed treatment effects. The introduction of pseudotherapy or control treatment conditions, corresponding to factors such as therapist contact, expectations of therapeutic gain, and credibility of procedures, enabled behavioral investigators to show that the specific treatment technique per se was responsible for therapeutic change. In addition, the favorable results obtained with this efficient technique appear to be lasting. Follow-up evaluations of the success of systematic desensitization, even with severe agoraphobics, have shown that treatment success was maintained after periods ranging from four to nine years (Marks, 1971; Munby and Johnston, 1980).

Research on systematic desensitization has not only focused on evaluation of its therapeutic efficacy but has also addressed the theoretical mechanisms that are responsible for its success. Briefly, Wolpe's original reciprocal inhibition hypothesis has been refuted and alternative explanations have become the subject of experimental analysis, as described below. What can be stated is that the necessary condition for successful reduction of most cases of phobic fear and avoidance has been identified: systematic, repeated exposure to fear-eliciting situations. Relaxation training and graded exposure via a hierarchy of stimulus scenes can facilitate exposure and thereby benefit treatment, but neither are critical ingredients for therapeutic success.

The explosion of research on process and outcome of systematic desensitization and related techniques of fear reduction that are discussed below requires comment. The quantity and methodological quality of this experimental research activity is unprecedented in the history of psychotherapy. Typically, in the psychological therapies, a new method has been introduced on the basis of anecdotal observations and uncritically accepted into clinical practice without subsequent experimental evaluation. That behavioral treatment methods for neurotic disorders, such as systematic desensitization, became the focus of intensive experimental investigation highlights the unique continuing link between experimental research in psychology and the clinical practice of behavior therapy.

There are a number of specific reasons for this upsurge of research on treatment methods. First, techniques like systematic desensitization were carefully specified and operationally defined so that they could be applied by different investigators in a standardized fashion. Replication of methods and results, a necessary feature of any scientific endeavor, therefore became possible. Contrast this development with the case of traditional psychotherapies, in which procedures are complex and poorly specified, making replication difficult, if not impossible. This view is shared by workers in the field of psychoanalysis, as demonstrated by the conclusions of Fisher and Greenberg, following their detailed review of psychoanalytic theory and practice (1977:411): "The field [of psychoanalytic therapy] is filled with vagueness, appeals to authority rather than evidence, lack of specificity in the definitions used, and unreliability in the application of techniques and dynamic conceptualization." Which, it must be added, makes evaluation well-nigh impossible.

A second reason for this research effort was the development of an experimental paradigm for studying fear reduction processes and procedures. Following well-established precepts from experimental psychology, behaviorally oriented investigators analyzed fear and phobic disorders in tightly controlled laboratory settings (so-called analogue research--see Figure 1) that permit precise specification of treatment procedures and detailed assessment of outcomes using objective measures of behavioral avoidance and physiological arousal (Lang, 1969). Laboratory or analogue studies of this kind enabled behavior therapists to examine theoretical mechanisms in treatment success and to refine specific techniques. Treatment methods were then extended to phobic patients in different clinical settings. The laboratory studies referred to above not only facilitated the development of more effective treatment methods but also led directly to major advances in our understanding of the basic nature of fear and its modification. To mention only one finding, it is now clear that fear is not a unitary phenomenon, but a complex concept of at least three loosely coupled components--overt avoidance behavior, physiological arousal, and verbal (subjective) responses. These three components of the fear response do not always correlate with each other; they may be differentially responsive to different treatment methods and can change at different rates. This information on the synchrony or lack of it of different measures of fear has given rise to several predictions about therapeutic change, suggests useful guidelines for assessing the nature of patients' phobic complaints, and provides a basis for new strategies for evaluating treatment outcome. The breakthrough in measurement and assessment in phobic disorders has had important effects on the assessment of a variety of other disorders, such as sexual dysfunction, pain, etc. (see Rachman and Wilson, 1980). The research on behavioral

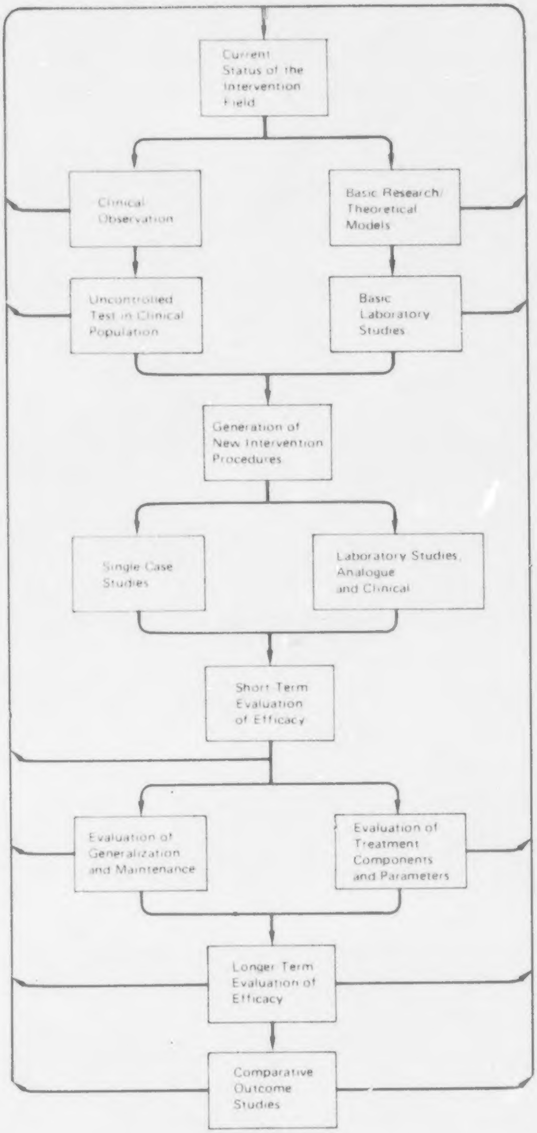
Current Status of the Intervention Field
Clinical Observation
Basic Research/ Theoretical Models
Uncontrolled Test in Clinical Population
Basic Laboratory Studies
Generation of New Intervention Procedures
Single Case Studies
Laboratory Studies, Analogue and Clinical
Short Term Evaluation of Efficacy
Evaluation of Generalization and Maintenance
Evaluation of Treatment Components and Parameters
Longer Term Evaluation of Efficacy
Comparative Outcome Studies

FIGURE 1 The flow of therapeutic research. SOURCE: Agras et al. (1979:106). Copyright © 1979 by W. H. Freeman and Company. Reprinted by permission.

methods of fear reduction represents a persuasive example of the mutually beneficial, reciprocal influence between basic research and clinical applications, a theme that is elaborated in a later section of this paper.

Flooding and In Vivo Exposure Unlike systematic desensitization, which relies on client-controlled, graduated exposure to anxiety-eliciting stimulus conditions, flooding involves therapist-controlled, prolonged exposure to high-intensity aversive stimulation without the soothing effects of relaxation training. Flooding may be conducted in imagination, but it is more usually done in vivo. The presentation of feared stimuli is expected to elicit a strong emotional response initially; however, continued exposure to these stimuli should result in a rapid decrease in fearful response. In order to ensure full exposure to the stimuli that elicit anxiety, the client is prevented from making an avoidance or escape response. Whether the exposure is presented in imagination or in real life, the client is strongly encouraged to continue to attend to the anxiety-eliciting stimuli despite the initial stressful effects this usually entails. For example, an agoraphobic client would typically be asked to imagine the experience of a sudden panic in a crowded supermarket, surrounded by an uncaring group of people, developing uncontrollable palpitations of the heart, being bathed in a nervous sweat, feeling intense shame and public embarrassment, fainting, and being taken by ambulance to a mental hospital. Such a scene would be presented continuously in imagination until the client's initial anxiety shows a definite decrease and the formerly frightening stimuli no longer elicit much distress. In vivo exposure is similar to flooding in vivo except that it may be conducted on a graduated or hierarchical basis. The purpose is to have clients confront their feared situations or objects without experiencing the intense levels of anxiety that are part of flooding.

Flooding and in vivo exposure methods derive directly from studies of animal conditioning and avoidance behavior (Levis and Hare, 1977). A typical study would be to place the animal (usually a rat or dog) in one compartment of an apparatus in which the subject could escape to a second compartment by making a response such as jumping over a barrier. A neutral stimulus (e.g., a tone) is presented, followed by an electric shock. Initially the animal manages to escape from the electric shock by making the response that gains access to the "safe" compartment. The rat learns rather quickly to anticipate and avoid the shock by moving into the safe compartment whenever the tone is presented. The tone becomes a conditioned fear stimulus, evoking fear that is reduced by the instrumental act of avoidance. One of the characteristics of this avoidance behavior that has attracted considerable theoretical and experimental attention is the great persistence the animals showed in continuing to make the avoidance response even though the shock was turned off. In some instances the animal would continue to avoid to the point of physical exhaustion, a phenomenon that was described as "pathological" (e.g., Solomon et al., 1953).

Behavior therapists seized on studies of avoidance behavior in laboratory animals as a convenient analogue to the study of phobic reactions. As noted above, Wolpe overcame avoidance behavior in his cats by encouraging exposure to the feared stimuli through graded presentations of the stimuli, eliciting a response that competed with or inhibited the conditioned fear reaction. Subsequently, more rapid methods of eliminating "neurotic" avoidance behavior were discovered. One such method involved simply preventing the animal from making the avoidance/escape response once the conditioned stimulus was presented by placing a barrier between the two compartments. Initially the animal would show strong fear reactions and struggle to escape. Eventually, however, the fear seemed to disappear and attempts to avoid or escape ceased. When the barrier was removed and the conditioned fear stimulus presented, the animal showed no avoidance behavior. The analogy to flooding and in vivo exposure treatment methods is obvious, except that phobic patients are not physically forced to confront their feared situations. Not only did this conditioning research on the elimination of long-lasting avoidance responses in laboratory animals suggest a specific treatment procedure, but it also helped to specify some of the important parameters that determine how well the technique works. For example, the research on the avoidance responses of animals showed that exposure to the conditioned fear stimuli should be continuous and protracted. Too brief a period of exposure would fail to eliminate the conditioned fear (or would even temporarily increase it), which might then motivate other forms of avoidance behavior. Precisely the same finding has been reported with phobic patients. To illustrate, one study demonstrated that two hours of continuous flooding in vivo was significantly more effective than four separate hours in one afternoon in the treatment of agoraphobic patients. Short exposures have produced a transient increase in fear.

Several well-controlled studies have demonstrated that flooding can be an extremely effective therapeutic technique in the treatment of different phobic disorders. Some of these studies have involved comparisons between flooding and systematic desensitization, the overall results indicating that flooding is more widely applicable, more efficient, and probably more effective (Marks, 1981).

Several lines of evidence clearly show that flooding and in vivo exposure are efficient and effective methods of treating phobic disorders. In order to go beyond relatively global, summary statements of numerous research investigations and to present a clearer picture of the nature of this evidence, I present the details of a specific project that was part of a clinical research program at Oxford University in England. It is a behavioral study that combines methodological rigor with clinical relevance.

Gelder and his associates (1973) compared flooding with systematic desensitization and a placebo control condition in the treatment of clinical phobias. In the placebo condition the therapist presented phobic images to initiate the clients' free association but made no attempt to control the content of subsequent imagery or verbal responses. The clients were told that this exploration of their feelings would enhance self-understanding and decrease their anxiety. All treatments were carried out by experienced therapists explicitly trained in the administration of the different methods. An attempt was made to induce a high expectancy of success in half the subjects by describing the treatment and the therapist chosen in very favorable terms, showing them a videotape of a client who had benefited from the treatment they were to receive. Treatment effects were evaluated in terms of measures of behavioral avoidance, blind psychiatric ratings, client self-ratings, physiological responsiveness, and standardized psychological tests. The adequacy of the control group in eliciting expectancies of treatment success comparable to those evoked by the two behavioral methods was assessed directly. Half the clients were agoraphobics, the other half a mixed group of specific phobias. Patients were assigned to treatments and therapists by means of a factorial design that permitted an analysis of the possible interactions among treatment effects, therapist differences, type of phobia, and levels of expectancy. Treatment duration was 15 weekly sessions. In sum, the Gelder et al. (1973) study was sufficiently well designed and well executed to answer the question of what treatment method had what specific effect on what problem in whom.

Both behavioral treatments, particularly flooding, produced greater improvement than the control condition on the behavioral avoidance tests, the physiological arousal measures, the psychiatric ratings of the main phobia, and patients' self-ratings of improvement. Simply put, flooding was roughly twice as effective as the powerful placebo control treatment. An important finding in this study was that the placebo control treatment was markedly less effective than both flooding and systematic desensitization with agoraphobics than with the other subjects. This result provides additional evidence that the success of behavioral methods such as flooding and systematic desensitization cannot be attributed solely to the role of placebo factors or expectations of favorable therapeutic outcome.

The experimental sophistication of studies of this kind, as noted above, in which a specific technique is compared with a highly credible placebo condition that controls for potential influences, such as the therapeutic relationship, expectations of therapeutic gain, and the like, represents a dramatic improvement over manifestly inadequate research on traditional psychotherapy methods (Kazdin and Wilson, 1978; Rachman and Wilson, 1980).

Another commendable aspect of this research by the Oxford group is their practice, regrettably rare both in the behavior therapy and the psychotherapy literature, of conducting long-term follow-ups. The six-month follow-up data showed that the treatment effects were maintained.

Subsequently, Munby and Johnston (1980) published a report of follow-ups of the agoraphobics from the Gelder study and those of two other related studies by the same group of investigators at Oxford five to nine years after therapy. Of the total of 66 agoraphobics who had been treated in these three studies, 95 percent were interviewed by a psychiatric research worker five to nine years later. Follow-up measures, repeating those used in the original studies, were compared with those obtained prior to treatment and six months after treatment ended. On most measures of agoraphobia the patients were much better at follow-up than they had been before treatment. The assessor's ratings suggested that there had been little change in the patients' agoraphobia since six months after treatment. Some of the patients' self-ratings showed evidence of a slight improvement over this period. No evidence of any untoward effects of this treatment was found. It should be noted, however, that a sizable number of these former patients had received additional treatment (behavior therapy or psychotropic drugs) over the follow-up interval, suggesting that caution be exercised in interpreting these long-term results.

Two other long-term follow-ups provide further support for the durable effects of in vivo exposure treatments of agoraphobic patients. In Holland, Emmelkamp and Kuipers (1979) followed up 70 outpatient agoraphobics, derived from a sample of 81 patients who had received exposure treatments four years previously. All information was obtained from questionnaires that were mailed to patients. Improvements in phobic fear and avoidance made during treatment were maintained, and on some of the measures further improvement occurred; there was also a reduction in depression in the follow-up period and no new neurotic disturbances developed. In Scotland, McPherson and his colleagues (1980), using a follow-up by mail of 56 agoraphobics who had shown improvement when treated with in vivo exposure, similarly found that treatment gains were maintained four years later.

Comparable therapeutic successes with flooding and in vivo exposure methods have been reported by different investigators in different countries. Moreover, aspects of the delivery of treatment have been systematically varied, for example, by comparing the results produced by individual and by group treatment. In vivo exposure conducted on a group basis is as effective as individual treatment, which is an important finding, since group treatment saves expensive therapist time and reduces the cost of treatment. Similarly, Marks and his colleagues in London described a program to train nurses to treat phobic and obsessive-compulsive disorders and examined the costs and benefits of such treatment (Marks et al., 1977). Nurse therapists obtained treatment results comparable to those obtained by psychologists and psychiatrists, with the advantage that the nurses required less training time. The benefits of treatment exceeded the costs when improvement continued for more than two years. This was largely the result of three factors: reduced use of treatment facilities after therapy, a reduction of absenteeism from work, and reduced family expenses as a result of enhanced functioning. Factors such as reduced suffering and increased enjoyment of life were not assigned a monetary value and can be regarded as added benefits.

To conclude, flooding and in vivo exposure methods have been shown to be effective in treating phobic disorders. The evidence indicates that they are the preferred psychological therapy methods for treating phobics. This conclusion does not imply that in vivo exposure methods should not, in some instances, be supplemented by other cognitive, behavioral, or pharmacological treatments, although the evidence is not clear on the value of these additional therapeutic procedures. For example, the behavior therapist might need to employ assertiveness training to overcome interpersonal problems, resolve marital conflicts, and ensure that family members assist in maintaining treatment-produced success. Problems of concurrent depression might require behavioral treatment directed primarily at depression or even the adjunctive use of antidepressant drugs. Complex problems require multifaceted treatment interventions.

Theoretical Mechanisms in Fear Reduction The common denominator in the demonstrable success of fear reduction methods like systematic desensitization, flooding, and in vivo exposure is the systematic, repeated exposure to feared situations. Yet this is only a description of what effective treatment minimally involves. The explanation for the success of exposure remains to be identified. The original theoretical account of these procedures does not fare well in the light of available evidence, although the specifics are beyond the scope of this paper (Wilson and O'Leary, 1980). (This example of a theory giving rise to an effective procedure, only to be subsequently proven untenable as an explanation of that procedure, is not uncommon in the psychological therapies.) A recent conference sponsored by the National Institute of Mental Health on behavior therapies led to the recommendation that the highest priority for future research be given to "studies that would answer basic process questions concerning the mechanism of action of exposure-based treatments in order to determine why these treatments work. It was concluded that answers to these questions would help to improve the effectiveness and efficiency of these treatments as well as making them more generally applicable" (Barlow and Wolfe, 1980). This recommendation from researchers and practitioners from this country and Europe indicates the importance they assign to the role that basic research has already played in the development of effective therapeutic methods and its potential role in refining and improving existing strategies. In the search for a better understanding of therapeutic fear reduction processes, current attention is focused primarily on two lines of experimental research.

The first is Bandura's (1977) laboratory research on self-efficacy theory. According to Bandura, treatment techniques such as desensitization, flooding, and variants of modeling are effective because they increase the client's expectations of personal efficacy. Efficacy expectations reflect a subjective estimate that one has the wherewithall to cope successfully with threatening situations. This mini-theory specifies the sources of information from which efficacy expectations derive and the major sources through which different modes of treatment operate. This cognitive theory of fear reduction methods generates different predictions from those of conditioning theory. Initial experimental tests of self-efficacy theory have yielded encouraging if not unequivocal results.

The point to be stressed is that self-efficacy theory was formulated and developed on the basis of carefully controlled laboratory experiments using subjects with specific phobias about snakes. The underlying assumptions of the theory can be traced to recent theorizing and research in cognitive and social psychology (e.g., information-processing and attribution theory) as well as experimental clinical analyses of fear reduction in behavior therapy. The predictions derived from this experimental research have led to important therapeutic applications with complex phobic disorders in clinical patients (e.g., agoraphobics), depression, and substance abuse. Regardless of the ultimate value of self-efficacy theory, it provides another example of the interplay between laboratory research and therapeutic applications that is fundamental to the development of an applied clinical science.

A second line of experimental research in fear reduction involves Lang's (1979) bioinformational theory of emotional imagery. The theory is derived from current cognitive conceptions about the nature of imagery and information-processing and psychophysiological analyses of fear arousal in laboratory subjects. In this propositional analysis, an image is seen as a conceptual network, the cognitive structure of which controls specific physiological responses and serves as a prototype for overt behavioral expression. The difference between this theory and previous conditioning theory analyses of imagery and conditioned autonomic responses (e.g., Wolpe's reciprocal inhibition theory) is obvious. Lang stresses that an image is not an internal stimulus to which a person responds (the stimulus + response conception of imagery in the behavior therapies). Rather, the person generates a conceptual structure that contains both stimulus and response propositions. Behavior change "depends not on simple exposure to fear stimuli, but on the generation of the relevant affective cognitive structure, the prototype for overt behavior, which is subsequently modified into a more functional form" (Lang, 1979:501).

One of the advantages of this theory is that it provides an explanation for the variable effects of exposure therapies in the treatment of phobics. Only those who generate the relevant affective stimulus and response propositions can be expected to respond successfully. Those who are unable to accomplish this primary processing of affective information will show poorer outcomes. Since Lang has been able to train subjects to improve their generation of affective response propositions, direct tests of this assumption of the theory are feasible.

Modeling Procedures Modeling refers to the process of observational or vicarious learning in which a person acquires new patterns of behavior or displays formerly inhibited actions as a result of watching a model perform these activities. The latter process, what Bandura (1977) calls the disinhibitory effects of modeling, are especially important in the treatment of fears and phobias. Modeling processes and procedures became increasingly important in behavior therapy during the 1970s. Major progress has been made in the scientific analysis of modeling during this period, both in terms of the underlying theoretical processes and applications to clinical and educational settings.

The previous sections on systematic desensitization and flooding described therapeutic techniques derived from studies of classical and instrumental conditioning of animals. The extensive use of modeling procedures in clinical and educational applications reflects the influence on behavior therapy of other areas of basic psychological theory and research--namely, social development and cognition. It was the emphasis on cognitive-social processes such as modeling, and the inevitable broadening of the conceptual and experimental base of behavior therapy, that helped to make Bandura's book, Principles of Behavior Modification (1969), such a decisive influence on the field. Acknowledging the demonstrable value of conditioning principles derived from animal research, he noted also their inherent limitations in conceptualizing the ways in which social behavior is acquired and modified. In order to account for social phenomena, it was necessary to modify existing learning principles and introduce new concepts that had been established through experimental studies of the acquisition and modification of human behavior in dyadic and group situations.

The chain of influence, whereby laboratory studies of modeling processes (particularly the research of Bandura and his associates at Stanford University) led to increasingly widespread clinical and educational applications of modeling procedures, is easily traced. As such, it provides a compelling example of the way in which basic psychological research on social development has improved our ability to ameliorate many human problems (Rosenthal and Bandura, 1978). Experimental research has established vicarious learning as a robust phenomenon in which children and adults learn new behavior by observing a model engage in a particular action. Contrary to conditioning principles, the observer need not emit the modeled behavior, nor is reinforcement necessary, in order for learning to take place. Moreover, we now know a great deal about the factors that influence the effectiveness of modeling and what the operative theoretical mechanisms are. In short, a model's effect is influenced by such characteristics as status and similarity to the observer, by the observer's ability to attend to, extract, and remember what he or she sees, and by procedural characteristics such as the consequences to the model, the number of models, and the range of situations in which the model appears.

Modeling procedures have been used extensively in therapeutic applications to develop new behavior or to overcome fears and phobias.[1] The latter practice is described below.

---

[1] Modeling has proved useful in changing a broad range of behavior, including verbal expressiveness, assertiveness, social competencies, problem solving, and self-

Bandura's initial studies demonstrated how modeling was effective in eliminating children's fears. Children with a fear of dogs watched a film in which they observed a model interact fearlessly with a dog. Assessment of these children's fears before and after exposure to the film showed that modeling was very effective in reducing fear and was significantly superior to simply exposing the children to a film of the same dog. The greater the number of models and the more varied the dogs with which they interacted, the greater the therapeutic effects. In other applications, extremely withdrawn preschool children increased their social activities after viewing a film in which an initially fearful peer model progressively interacted more frequently with other children. Similarly, the use of filmed models can significantly reduce children's fears of dental treatment as well as reduce the anxiety associated with undergoing surgery. The basic method can be understood from a brief description of one of the studies in this area. Melamed and Siegel (1980) prepared a film that depicted the experiences of a seven-year-old boy undergoing a hernia operation. The child described his own feelings, the fears that he experienced at each stage, and his resolution of them. The observer viewed the child's progress through the admission process, ward orientation, examinations by the surgeon and anesthesiologist, return to the recovery room, reunion with his parents, and hospital discharge. Upon hospital admission, 30 children about to undergo surgery for hernias, tonsillectomies, or urinary problems were shown either the modeling film or a control film about a boy's trip in the country. In addition, all children received preoperative preparation, which included demonstrations and explanations of the surgery and recovery process by a social worker and a visit from the surgeon, who again explained the surgery to the child and his or her parents. The children's self-reports of anxiety, staff observations of

---

control strategies. The educational and preventive possibilities of modeling have begun to be analyzed. Early laboratory studies by Bandura showed how exposure to aggressive models produced significant increases in aggression in children. Evidence from many sources indicates that aggressive models on TV lead to greater aggression in viewers. The portraval of women's roles and the adverse effects of "junk food" commercials on children's nutritional preferences have also been studied.

their behavior, and physiological measures of anxiety arousal all showed greater reduction of anxiety both at preoperative (night before operation) and postoperative (examination three to four weeks after surgery) assessments for the group who were exposed to the modeling film. Parenthetically, it should be noted that this example of behavioral procedures being incorporated within medical treatment settings and services is only one aspect of a much wider trend. The increasing utilization of behavioral interventions within medicine, with respect to both prevention and treatment, is known as behavioral medicine. (This development is discussed more fully in the paper by Krantz et al., in this volume.)

With adult phobics, modeling has been found to be either equal or superior to systematic desensitization as a fear reduction technique. In addition to specific phobias, such as fears of small animals or heights, modeling has proved effective in treating anxiety that results in sexual dysfunction. For example, women with both primary and secondary orgasmic dysfunction (the inability to achieve orgasm and experience sexual satisfaction) were successfully treated by having them observe videotaped vignettes showing graduated sexual behavior with instructions to practice these activities at home with their spouses. One of the obvious advantages of videotaped modeling methods such as those referred to above is their cost-effectiveness and their potential for expanding health and mental health service delivery systems (Rosenthal and Bandura, 1978).

Participant modeling is a method in which the client first observes an appropriate model's actions, then is asked to engage in the feared behavior with active guidance, support, and corrective feedback from the therapist. This procedure is clearly more effective than modeling alone or imaginal systematic desensitization. Phobics who are not helped by symbolic modeling or imaginal desensitization are often readily aided by participant modeling, a procedure that closely resembles the therapist-assisted in vivo exposure methods discussed in the previous section. Whether or not modeling increases the efficacy of in vivo exposure will depend on the type and severity of the problem in particular patients. Rosenthal and Bandura offer the practical observation that "treatment planning is constrained by the specific activities the client will or will not undertake. Clients attempting or refusing to perform a given task sets the limits on the momentary content of treatment" (p. 640). Participant modeling includes a number of performance aids that are designed to prompt the recalcitrant client to make contact with the phobic situation. Among these aids, the verbal support of the therapist and prior modeling might prove to be decisive in cases of extreme fear.

The roots of modeling procedures in cognitive-social learning theory are evident from Bandura's assumption that the source or mode of the disinhibiting information that forms the basis of modeling is less important than the ultimate impact the information has on the cognitive mediating processes that are assumed to regulate behavior. Thus, in addition to live and filmed models, the effects of covert modeling, in which the person imagines a model confronting phobic situations, have been evaluated. In a series of well-controlled studies, taking the cue from basic psychological research and therapy, Kazdin and Smith (1979) demonstrated that covert modeling is more effective than control conditions in reducing phobias and increasing assertive responses in interpersonally inhibited individuals. Several procedural factors that critically influence treatment outcome have been identified. The effects of covert modeling are enhanced if multiple rather than a single model is used and if their behavior is followed by reinforcing consequences. If individuals imagine a "coping" model who gradually overcomes fear as opposed to a "mastery" model who performs fearlessly from the outset, greater behavioral change ensues. These findings are reassuringly consistent with the underlying theory and with similar results obtained with live and symbolic modeling procedures. Although not as powerful as participant modeling, and an unsuitable method for many patients for one or more reasons, the practical advantages of covert modeling are obvious since imagery provides an efficient means of rehearsing phobic scenes or inhibited activities that would be difficult if not impossible to arrange on an actual behavioral basis. This was one of the reasons that led Wolpe to the use of imagery in systematic desensitization. The limits of covert modeling methods require definition through more research.

Obsessive-Compulsive Disorders

There is wide agreement among clinicians of different theoretical orientations that obsessive-compulsive disorders are among the most severe and disabling of psychiatric problems. They have remained resistant to various psychological and somatic therapies, thereby providing a testing ground for potentially effective methods.

As in other areas, the behavioral treatment literature shows a definite progression toward the development of increasingly refined and more effective therapeutic techniques. Our recently improved capacity for modifying these powerful and resistant problems can be attributed mainly to two developments. First, there has been a switch from imaginal to in vivo treatments, which appear to make a powerful contribution to the treatment of obsessional disorders. Second, response prevention methods were introduced in the late 1960s (Rachman and Hodgson, 1980). Flooding in vivo involves directing the patient (e.g., an obsessive-compulsive hand-washer with a fear of contamination) to make repeated contact with the most fear-provoking situations (e.g., supposedly contaminated articles) as soon as possible. As in the case of phobic disorders, this in vivo exposure may also be graduated. In response prevention the patient is either urged to refrain from engaging in compulsive rituals or, in the case of inpatient treatment, is provided with continuous nursing supervision to ensure that compulsive acts are prevented. The logic of this method derives directly from the animal conditioning research discussed above. Persistent avoidance behavior in rats or dogs is rapidly eliminated by the simple expedient of physically preventing the avoidance behavior in response to the conditioned fear stimulus. Compulsive rituals are viewed, analogously, as avoidance responses that can be eradicated in the same manner.

The nature and effects of response prevention can be illustrated with reference to the following study of five hospitalized, severely disturbed obsessive-compulsive patients (Mills et al., 1973). A noteworthy feature of this study is the use of a type of experimental methodology that permits evaluation of the efficacy of specific treatment techniques. Each patient served as his or her own control in evaluating the effect of sequentially introducing, then removing, a therapeutic procedure. Evaluations were made in terms of individual compulsive behavior, which was reliably recorded on a continuous 24-hour basis. The outcome of the treatment of one of the clients, a compulsive hand-washer, illustrates the general pattern of results.

Recordings of the frequency of hand-washing and urges to engage in this behavior were made over the first eight days of the baseline period. During the last eight days the client was exposed to objects that typically elicited the urge to hand-washing. Predictably, hand-washing increased at this point. Next, a placebo condition was introduced in which the client was given two "drugs" (actually a glucose capsule, four times a day, and an injection of saline, one daily) that she was told would enable her to control her hand-washing compulsion. Hand-washing increased slightly during this phase to an average of 60 episodes a day. With the introduction of response prevention during the third phase of this experiment, no hand-washing was possible; the handles were removed from the sink and the shower in the client's room. During phase four the client was once more able to wash if she so desired. However, episodes of hand-washing were minimal, and this improvement was maintained during the final baseline phase of the study, when no treatment intervention was in effect. Note, however, that the client still reported urges to wash. The difference was that now, as a result of treatment, she could control her behavior. Additional exposure treatment in her home setting, following release from the hospital, resulted in a reduction in the urge to wash her hands compulsively. Similar substantial and lasting improvements were achieved with all four of the other subjects.

The available evidence, despite its inevitable gaps and inadequacies, supports the view that behavior therapy is capable of producing significant modifications in these disorders. The relevant data consist of uncontrolled clinical trials and controlled outcome studies carried out in this country, Europe, and Australia. The quality of the controlled studies of these disorders does not match that of research on phobic disorders, suggesting caution in interpreting the available data. Summarizing this evidence, Rachman and Wilson (1980) concluded that:

> Behavioral treatment produces significant changes in obsessional problems, and rapidly at that. Clinically valuable reductions in the frequency and intensity of compulsive behavior have been observed directly and indirectly. Significant reductions in distress and discomfort are usual, and psychophysiological changes of a kind observed during the successful treatment of phobias, also occur. The admittedly insufficient evidence on the durability of the induced change is not discouraging; allowing for the provision of booster treatments as needed, . . . the therapeutic improvements are stable. The

> successful modification of the main obsessional problems often is followed by improvements in social and vocational adjustment. In all the series and controlled studies reported so far, some clear failures occurred--the failure rate ranges between 10% and 30%. The reasons for such failures are not known. . . . The influence of depression on obsessional disorders is acknowledged to be of importance . . . but the presence of depression is not necessarily an obstacle to treatment.

As in the case of complex agoraphobic disorders, the use of in vivo exposure and response prevention methods with obsessive-compulsives must often be supplemented by other behavioral or pharmacological methods (e.g., antidepressant drugs to change depressed mood; see Marks et al., 1980).

Psychotic Disorders

The application of behavior therapy to psychotic disorders has centered almost exclusively on the treatment of chronic mental patients in psychiatric institutions. The vast majority of these mental patients have been diagnosed as schizophrenic, and treatment interventions have been confined largely to the use of the principles and procedures of operant conditioning. The product of Skinner's seminal research and writings, operant conditioning comprises a set of philosophical assumptions about behavior, a distinctive methodology for studying behavior, and a number of learning principles that have been widely used to modify diverse kinds of behavior. Nowhere in behavior therapy is the link between underlying theory and experimental research, and subsequent applications to human problems, clearer than in the case of the operant-conditioning approach. Operant-conditioning principles and procedures were all formulated and developed through the experimental analysis of the behavior of laboratory animals well before they were extended systematically to children and mental hospital patients circa 1960.

Philosophically, Skinner's operant-conditioning position is one of radical behaviorism, in which overt behavior is regarded as the only acceptable subject of scientific investigation. Subjective processes or private events, such as thoughts or feelings, are viewed as epiphenomena: They can never have a causal impact on behavior but are seen as correlates of behavior that are a function of environmental influences. However, it is probably accurate to assert that most behavior therapists reject this narrow conceptualization of behavioral control and hold broader views of psychological functioning. Of course, operant-conditioning procedures can and frequently are employed without acceptance of the tenets of radical behaviorism.

A significant contribution of the operant-conditioning approach has been the development of a methodology for evaluating the effects of treatment in single cases. Repeated, objective measurement of an individual's behavior under controlled conditions, in which therapeutic change is evaluated relative to the person's own performance, is the hallmark of operant methodology. Entailing a variety of different single-case experimental designs, this methodology permits the investigator to show that behavior change in an individual client is the result of specific treatment interventions and not simply due to the passage of time, placebo reaction, or some other uncontrolled event (Hersen and Barlow, 1976). The Mills et al. (1973) study of the treatment of obsessive-compulsive patients (referred to above) is an example of this experimental methodology. The introduction of these single-case experimental designs has significantly improved our ability to evaluate the effects of different treatment methods under conditions in which conventional, between-group designs and statistical analyses are unsuitable. They occupy an important place in the overall schema for evaluating therapeutic outcome, as shown in Figure 1.

As a set of principles for modifying behavior, the operant-conditioning approach has been particularly influential. Operant conditioning emphasizes the relationship between observable behavior and its environmental consequences. Behavior change is said to occur when certain consequences are contingent on the occurrence of behavior. Thus the cornerstone of operant-conditioning approaches, the principle of positive reinforcement, refers to an increase in the frequency of behavior that is followed, contingently, by a rewarding environmental event (i.e., a positive reinforcer). For example, if one allows a pigeon to peck a lighted key, arranges for each peck to be rewarded with a grain of food and the absence of pecking to go unrewarded, the rate of pecking will increase. Similarly, if one allows a hospitalized psychotic patient to gain access to a highly desired activity (e.g., watching TV) contingent on the performance of a specified act (e.g., to begin to wash and groom himself without nursing care), the latter behavior will reliably increase in frequency.

It is often the case that a new response cannot be established directly by reinforcing because the response may never occur. The desired behavior may be so complex that the elements that make up the response are not in the repertoire of the individual. Hence the critical element in operant conditioning is shaping, a procedure in which the terminal behavior is achieved by reinforcing small steps or approximations toward the final response rather than reinforcing the final response itself. Punishment is the presentation of an aversive event or the removal of a positive event contingent on a response that results in a decrease in the frequency of that response. If a painful electric shock is contingent on a rat's pressing a lever that once resulted in positive reinforcement, the lever-pressing will quickly cease. If one systematically withdraws a particular highly desired privilege each time a mental hospital patient displays violent behavior, that violence will diminish. Armed with these and several other learning principles, behavior therapists have undertaken to modify a remarkably broad array of problems across widely varying populations.[2]

In the systematic progression from experimental research to general clinical applications, a theme that is repeated throughout this paper, the initial extension of operant-conditioning principles to chronic patients in mental hospitals was strictly limited to studies designed to show behavioral change and to relate it to specific manipulations of reinforcement contingencies. Among the circumscribed responses that were modified in these single-case experimental designs were basic self-care activities (e.g., washing, grooming, and so on), attendance at various therapeutic activities, social withdrawal, and mutism or bizarre verbalizations. The clinical significance of changes in these limited behavioral targets was not of primary concern, and the effects, while confirming the generalizability of operant-conditioning principles from animal research to the modification of human behavior, fell far short of satisfactory therapeutic outcomes.

---

[2] Operant techniques have been applied, with uneven success, to the treatment of psychiatric patients; mentally retarded persons; children with disorders such as autism, conduct problems, and hyperactivity; predelinquents and delinquents; drug addicts and alcoholics; people with eating disorders; and those with different neurotic complaints. In reviewing these diverse applications Kazdin (1978b) also notes that "a major development is the application of operant techniques to alter community relevant behaviours including energy conservation, littering, use of mass transit, recycling of waste material, job procurement . . . and community self-help behaviours" (p. 561).

Successful demonstrations that the behavior of mental hospital patients could be modified through the creative use of operant conditioning led to more ambitious interventions with explicit therapeutic objectives. Of paramount importance in this effort was the development of the token economy, a behavioral program designed to produce change in entire social systems or groups of patients (e.g., a psychiatric ward of a hospital). Briefly, the token economy consists of the following main elements: instructions about what behavior will be reinforced (the target behavior); precise definition of that target behavior; back-up reinforcers, which are the "good things in life" or what people are willing to work for (e.g., access to desired activities); the tokens that represent the back-up reinforcers (e.g., a plastic chip or a numerical rating); rules of exchange that specify the number of tokens required to obtain the back-up reinforcers. Tokens are generalized conditioned reinforcers, which have a number of advantages: They bridge the gap between the target behavior and the back-up reinforcers, they permit the reinforcement of any response at any time, and they provide the same reinforcement for different patients who have diverse preferences in back-up reinforcers.

In a pioneering study in this area, Ayllon and Azrin (1965) showed that severely disturbed schizophrenic patients could be encouraged to engage in social and vocational activities that they had failed to display throughout their hospitalization. On a rudimentary level they successfully acquired behaviors that would ultimately help them to function more effectively outside the hospital--for example, housekeeping, following instructions, and interacting with fellow employees.

Subsequent studies attempted modification of more complex, clinically significant behavior, such as independent problem-solving and decision-making skills. In order to cope with the problem of relapse and readmission to the hospital following discharge due to therapeutic improvement as inpatients, explicit after-care was devised to integrate discharged patients into the communities to which they returned on a more permanent basis. A study by Fairweather et al. (1969) illustrates the use of the principle of fading, or gradual withdrawal of reinforcement provided by the treatment program, to the point at which the natural reinforcing contingencies in the patients' home communities sustained their improved functioning. Once patients were functioning adequately within the hospital, they were transferred to a semiautonomous lodge located in the community. Under the initial supervision of a hospital staff member and subsequently of a lay person, the patients assumed full responsibility for running the lodge, including the purchase of food and the preparation of meals, regulating the administration of medication, managing their own financial affairs, and operating a money-making janitorial service. The necessity of specific training in relevant job skills is underscored by the failure of these patients to perform even simple tasks, such as gardening, until they were deliberately trained. The operation of the lodge became fully autonomous after 33 months of functioning: All income that was earned was distributed among the lodge members according to each person's productivity and role in the system. The control group received the traditional assistance and outpatient therapy that hospital patients get after being discharged. The superiority of the lodge members over the control group was most evident in terms of the amount of time they maintained themselves in the community over a 40-month follow-up and the percentage of time that they were gainfully employed during this period.

A number of studies have compared the token economy to other forms of treatment in the care of psychotics and other patients of mental hospitals. For convenience these studies can be grouped according to whether the comparison treatment consisted of routine custodial ward treatment or a specific alternative form of therapy. Typically, routine ward treatment involves custodial care, although several unspecified activities and a small amount of therapy of some sort may be included. Alternative treatments refer to active procedures that are more well specified than custodial ward care and are associated with an explicit rationale from which the procedures are derived.

The results of eight studies comparing the token economy to routine hospital treatment has been summarized by Kazdin and Wilson (1978:75):

> Comparisons of reinforcement programs with routine ward care indicate a relatively consistent pattern. Reinforcement programs have led to improvements on measures of cognitive, affective, and social aspects of psychotic behavior, on specific behaviors in interviews and on the ward, and in global measures of adjustment or discharge from the hospital and subsequent readmission. Patient improvements in these areas are much greater for reinforcement than for routine care wards. Regrettably, comparisons of reinforcement and routine ward programs have been beset with methodological problems, including bias in subject selection and assignment, confounds of treatment with changes in the physical ward and in the hospital staff, and ancillary features of the hospital environment.

Consequently, although suggestive of the superiority of token economies, the available evidence falls short of unequivocal confirmation.

Kazdin and Wilson reported seven studies comparing the token economy to alternative forms of therapy in the treatment of psychotic patients. As a whole these studies were superior in methodological quality to those in which the comparison condition was routine hospital care. The specific treatments to which the token economy has been compared include role-playing, verbal psychotherapy, play therapy, recreational therapy, and milieu therapy. It should be noted that the verbal psychotherapy treatments in three of the seven studies reviewed were quite different from one another and should not be interpreted collectively as a test of psychotherapy. Kazdin and Wilson concluded that in "almost all of the available studies, reinforcement techniques have been more effective than the comparative treatment" (p. 81). Here too, however, methodological problems preclude any firm conclusion about the superiority of behavior therapy, encouraging as the data are.

Among the shortcomings of this literature have been the focus on limited ranges of patient behavior and the absence of systematic follow-up evaluations. A study that overcomes these shortcomings and provides information on the limits of change that can be achieved by severely disturbed patients within the staffing restrictions of public mental hospitals in this country has been reported by Paul and Lentz (1977). The study is destined to become a landmark in the care of the chronic mental patient. In terms of its methodological excellence, its attention to detail, the specification of the treatment methods, staff training and administration, the comprehensive nature of its measurement of outcome, and its broad evaluation of cost-effectiveness, it is unprecedented.

The subjects were chronic mental patients, all of whom had been diagnosed as schizophrenics, were of low socioeconomic status, had been confined to a mental hospital for an average of 17 years, and had been treated previously with drugs and other methods without success. So minimal was their level of self-care and so pronounced the severity of their bizarre behavior that they had all been rejected for any sort of community placement. Paul and Lentz described these subjects as "the most severely debilitated chronically institutionalized adults ever subjected to systematic study."

Of these subjects 28 were assigned to each of three treatment groups so that the groups were "identical on level and nature of functioning and on every characteristic of potential importance to treatment responsiveness." Two identical adjacent units were established at a mental health center to house two psychosocial programs--one based on milieu therapy and the other based on social learning therapy. Both were staffed by the same personnel at a level equal to that of staff at a comparison state hospital. The third group received typical hospital treatment at the comparison state hospital. As patients were released into the community, they were replaced by similar subjects from the original pool of patients.

The social learning treatment consisted of the direct application of experimentally established principles of learning, including classical conditioning, a variety of reinforcement procedures, and the token economy. The milieu therapy consisted of creating a therapeutic community structure in the institution. Within this therapeutic community the focus was on the communication of positive expectations, group cohesiveness and group pressure directed toward normal functioning, and group problem solving, in which the residents (as the subjects in this study were called) were treated as responsible people rather than as custodial cases. The routine hospital therapy consisted of typical state hospital treatment of the chronic schizophrenic patient, emphasizing chemotherapy, custodial care with practically no psychological treatment, and little positive expectation of patient improvement. The social learning and milieu treatments were "equally high-prestige programs in identical physical settings with exact equation in the degree of operationalization, clarity, specificity, explicitness, and order provided for both staff and residents. Both programs also provided identical activity structure and focus upon specific classes of behavior, with the same staff not only conducting both programs, but equating time and focus within programs, with both running concurrently over the same time periods, subject to the same extraneous events" (Paul and Lentz, 1977:423). As a result of this unprecedented experimental control, conclusions about comparative efficacy of the different treatments can be drawn. The social learning treatment showed consistent superiority over both the milieu and routine hospital treatments. Behavioral observation of patients' functioning in the hospital setting across the four and a half years the programs were in effect indicates that both the social learning and the milieu therapy programs led to major improvement at the end of treatment.

The social learning treatment produced significantly greater reductions in bizarre behavior and increases in adaptive behavior, such as self-care and interpersonal skills, than milieu therapy. Assessment of overall functioning on several standardized rating scales showed that the social learning treatment resulted in significantly greater improvement than the milieu therapy at each six-month evaluation and the routine hospital treatment at all but two of these different assessment periods. The effects of milieu therapy on these scales were less apparent than on the continuous behavioral observations. Patients in the routine hospital treatment failed to show a significant change in functioning over the four and a half years of the program.

The social learning treatment produced significantly greater release rates than either the milieu therapy or the routine hospital treatment. In turn, milieu therapy was superior to the routine hospital program, but the majority of releases were to private boarding and care homes (88.9 percent, 84.2 percent, and 100 percent, respectively, for the social learning, milieu therapy, and hospital programs) as opposed to completely independent functioning in the outside world. Of the original subjects, only 10.7 percent of the social learning residents and 7.1 percent of the milieu therapy residents achieved release to independent functioning and self-support without rehospitalization. None of the patients from the routine hospital program was successful in this respect. Paul and Lentz themselves note that the level of functioning required for remaining in extended-care facilities in the community was "marginal." Nevertheless, given the severity of the subjects' problems and their chronically low level of functioning prior to treatment, successful release to community facilities such as boarding homes was a notable accomplishment.

A follow-up after one and a half years revealed that the social learning program maintained its clear superiority over the other two groups, with well over 90 percent of the residents remaining continuously in the community at the time of the final follow-up. Over 70 percent of the residents treated by the milieu therapy program were still in the community at the final follow-up, compared with fewer than 50 percent of the patients from the hospital group. Not only was the social learning treatment the most effective, but it was also the least expensive--based on all the data from objective and standardized rating measures during the four and a half years that the treatment programs were in place, release rates and community stays, and cost-effectiveness figures. Paul and Lentz concluded that "social learning procedures clearly emerge as the treatment of choice for the severely debilitated chronically institutionalized mental patient" (p. 389). If there is any criticism of this study, it would be that a longer follow-up, perhaps on the order of five years, would be desirable. A major index of the success of the treatment of the chronic mental patient is the degree to which such a patient is reintegrated into the community.

In all, these findings support the view that effective and feasible procedures can be used to treat the chronic mental patient. Aside from its implications for the care of the chronic mental patient, the Paul and Lentz study also serves as a methodological model of comparative outcome research with a severely disturbed population.

Self-Control: Principles and Applications

Behavior therapy procedures that were derived from experimental studies with animals emphasize environmental or external influences on behavior (e.g., the token economy). Increasingly, however, the role of self-regulatory processes in behavior change has become a primary focus of experimental research and practical application in behavior therapy. In processes of self-control, individuals initiate attempts to change their own behavior in a manner that is relatively independent of external influences. They set their own goals, monitor and evaluate their own performance, and act as their own reinforcing agents.

Dating from the 1960s, the laboratory research and innovative theorizing of Bandura (1969) and Kanfer (Kanfer and Phillips, 1970) were the seminal influences on the development of self-control principles and treatment procedures. Using nonclinical populations of children and college students as subjects and adopting somewhat different experimental paradigms, these two investigators and their colleagues completed research that yielded a set of consistent findings on self-control processes. Self-control is an active area of theorizing and research, and recent studies have served to amplify, modify, and refine existing principles. Some of these laboratory findings on self-control processes that have proved effective in applied settings can be briefly summarized in terms of the various components of the social learning model of the self-regulation of behavior.

Self-Monitoring and Recording

Accurate observation (monitoring) and recording of precisely defined behaviors that are the target of change are fundamental to behavioral self-control procedures. Self-monitoring is not only the basis for subsequent self-evaluation and self-reinforcement, but it can also produce behavior change directly itself. This potential reactivity of self-monitoring has been shown to be a function of the specific properties of the behavior in question as well as when and where it is self-monitored.

Standard Setting and Self-Evaluation

Individuals set standards against which they evaluate their performance. These standards are the product of differential reinforcement by parents and other societal agents. Processes of social comparison become involved because in the case of most performances objective criteria of adequacy are lacking: Hence the attainments of other persons must be used as the norm against which meaningful self-evaluation can be made. Standards are also acquired and maintained vicariously. Studies show that subjects who have been exposed to models setting low standards tend to be highly self-rewarding and selfapproving for comparatively mediocre performances; by contrast, persons who have observed models adhere to stringent performance demands display considerable self-denial and dissatisfaction for objectively identical achievements. Personal standards or goals do not automatically activate self-evaluative processes that then help to determine behavior. Among the properties of goals that increase the likelihood of self-evaluative reactions are their specificity, the level at which they are set, how realistic they are, and whether they are relatively immediate or distant.

Self-evaluation training, combined with rewards, has helped to improve children's academic and social behavior. For example, when children with severe emotional problems were taught to evaluate their classroom behavior on a 1-10 rating scale and were then given feedback and rewards for appropriate evaluations and behavior, they were able to maintain the behavioral change that was effected through a token reinforcement program. Furthermore, it appears that once accurate self-evaluation skills have been taught, changes in academic and social behavior may be maintained even after a token program is removed. Presumably the children adopt some of the values that are being taught in the self-evaluation program (Turkewitz et al., 1975).

Self-Reinforcement

In process of self-reinforcement, individuals have access to freely available rewards. Yet they make the self-administration of any reward conditional on performance that matches or exceeds a self-selected standard. Laboratory studies have shown that self-reinforcement can maintain behavior as effectively as reinforcement that is externally administered. The basic research design involves one individual or group of individuals who select their own standards for reinforcement and another individual or group of individuals who have standards of reinforcement determined by the experimenter. An individual in the self-selection group decides how hard he or she wishes to work (for example, how many problems he or she wishes to complete) before receiving a reinforcer. A second individual is later given the same standard for reinforcement as was chosen by the first. The standards that are externally imposed are yoked or matched to the self-selected standards to ensure comparability. Studies in applied settings indicate that both self-imposed and externally imposed standards of reinforcement increase selected performance, but the two methods do not differ from each other. On some laboratory tasks, self-determined reinforcement contingencies have been shown to produce greater resistance to extinction than externally determined contingencies. Children exposed to both self-imposed and externally imposed reinforcement contingencies prefer self-determined contingencies.

Self-evaluative and externally occurring consequences may conflict, for example, when certain courses of action are approved and encouraged by others, but if carried out would give rise to self-critical and negative self-evaluative reactions. Under these circumstances the effects of self-reinforcement may prevail over external influences. Conversely, behavior may be effectively maintained by self-reinforcement operations under conditions of minimal external support, lending consistency and stability to actions in the face of rapidly changing social environments (Bandura, 1977).

Self-Instructions

Self-instructions refer to the self-talk that takes the form of prompts, guides, or demands. Explicit self-instructional training can significantly influence the behavior of both children and adults in laboratory studies (e.g., resistance to temptation and impulsiveness) (Meichenbaum and Asarnow, 1979). The training, which was heavily influenced by concepts from developmental psychology of the internalization of verbal control of behavior, is illustrated in the typical program for teaching impulsive children to modify their nonverbal behavior. The program includes: (1) an adult modeling a task while talking to himself out loud; (2) the child performing the task under direction of the model (guidance from adult); and (3) the child performing the task while instructing himself aloud, and finally covertly. In applied settings (e.g., the treatment of hyperactive children), self-instructional training appears most useful as part of a broader behavioral program.

Singly, collectively, and in combination with other behavioral methods, the self-control principles described above have been extensively applied to a wide range of clinical disorders and educational problems. Consider the problem of obesity, a condition that has remained resistant to lasting modification by other psychological and even pharmacological methods and that has been the target of intensive treatment with behavioral self-control strategies. Self-monitoring has proved invaluable not only as a method of behavioral assessment--namely, of identifying the conditions under which problems occur--but also as a technique of change per se. For example, obese patients taught to self-monitor daily caloric intake show reliable weight losses. Greater weight losses are found if patients set realistic, short-term weight reduction goals. When systematic self-reinforcement is added to goal-setting and self-monitoring, weight loss is superior still. Stimulus control is a principle of operant conditioning that is often used in conjunction with self-control principles. In the treatment of obesity the patient is instructed to limit eating to specific situations and times and not to associate eating with other (distracting) activities such as reading or watching TV. When combined with other learning strategies behavioral self-control programs have been shown to be the preferred approach for mild to moderate cases of obesity (Wilson and Brownell, 1980).

Other addictive disorders, including cigarette smoking and problem drinking, have been modified through the use of behavioral self-control strategies alone or in conjunction with other behavior therapy techniques (Stuart, 1977). The thrust of the applications of self-control strategies discussed thus far has been treatment of existing disorders. The same principles and procedures also hold considerable promise for prevention of serious behavioral and health problems. For example, the Stanford Heart Disease Prevention Program, a three-community field study that was designed to reduce the incidence of cardiovascular disease by modifying risk-producing attitudes and behavior (e.g., cigarette smoking, high cholesterol intake, overweight, and so on), drew heavily on social learning theory in general and self-control principles in particular (Maccoby et al., 1977). The success of this program encourages the view that self-control strategies can be communicated effectively in large-scale preventive efforts in the field of public health.

Stress-related problems have been a major focus of cognitive and behavioral self-control strategies. Among the specific problems that have been treated successfully with methods such as self-instructional training are social anxieties, lack of assertiveness, severe anger disorders, and tension headaches (Kendall and Hollon, 1979).

Biofeedback

In the late 1960s Miller caught the attention of the world of psychobiology by demonstrating that the visceral responses of rats could be directly modified through operant-conditioning procedures (Miller, 1969). Prior to Miller's research it had been believed that instrumental learning applied only to skeletal responses. Visceral responses (e.g., changes in heart rate) that are functions of the autonomic nervous system were presumed to be modified only by more "primitive" associative learning via classical conditioning methods. The demonstration of instrumental control of visceral responses had a dramatic impact on basic and applied research, even though subsequent animal research revealed progressively smaller effects of reinforcement contingencies on autonomic responses. Attempts were promptly made to modify similar physiological functioning in people by providing them with moment-to-moment feedback (in the form of lights, tones, or graphic displays) of their response patterns. The idea was to increase a person's self-control over his or her physiological functioning, to bring under voluntary control bodily functioning that had hitherto been viewed as involuntary in nature. The biobehavioral specialty area of biofeedback was born.

Hypertension, cardiac abnormalities, headaches, migraine, and several other disorders have become the targets of therapeutic change through these learning-based procedures. Exaggerated claims have obscured the effects of biofeedback. An objective evaluation of the value of biofeedback shows that its therapeutic effects are real but limited. For example, in the treatment of migraine and tension headache, biofeedback appears to be significantly more effective than either medical or psychological placebo treatments. However, it is no more effective than progressive relaxation training. A similar verdict holds true for the treatment of hypertension and other psychophysiological disorders (Silver and Blanchard, 1978). Although biofeedback procedures have promise in other areas (e.g., a vascular disorder such as Raynaud's disease), the appropriate controlled outcome research remains to be completed. Beyond these therapeutic applications, biofeedback is proving to be an invaluable scientific research tool that facilitates the detailed analysis of the functional role of physiological variables in behavior.

The therapeutic potential of progressive relaxation training was rediscovered by Wolpe, who made it a key component of his systematic desensitization technique, described in an earlier section. Although relaxation training is not essential for the success of behavioral treatment methods for phobic disorders, it has proven to be effective in the treatment of other stress-related problems. In these applications relaxation training is taught as a self-control skill that the patient learns to implement as necessary. For instance, relaxation training produces significantly greater reductions in blood pressure than placebo or other control procedures. These reductions in blood pressure are clinically significant and may compare favorably with those produced by antihypertensive medication in some patients. The full magnitude and generalizability of relaxation training as a treatment of hypertension are still unknown. At a minimum, however, relaxation is a very useful adjunct to medication in the treatment of hypertension in patients whose blood pressure remains high despite medication (Jacob et al., 1977).

Other series of controlled studies have shown that progressive relaxation training, as a self-control procedure that the patient practices actively and regularly, can eliminate tension headaches and overcome some forms of insomnia. Continuing experimental research has begun to isolate the critical ingredients in this procedure, indicating that its clinical value can be refined. Cost-effective and readily disseminable, relaxation training offers a simple yet effective means of coping with many stress-related problems.

FROM RESEARCH TO CLINICAL PRACTICE: AN INTEGRATIVE FRAMEWORK

Figure 1 shows the different levels of analysis along the continuum from basic research to clinical practice in the development of effective treatment methods. Ultimately, neglect of any of these stages in the flow of therapeutic research and the complex interrelationships among them will undermine progress toward a scientifically based, clinically tested treatment approach.

The progression of therapy research within this framework can be illustrated by referring back to the earlier description of behavioral methods for the treatment of anxiety disorders. First, Wolpe derived a new technique, desensitization, from basic animal conditioning research.

Next, it was tested in controlled laboratory (analogue) studies using carefully selected subjects with homogeneous phobias (e.g., Lang's 1969 research) and in single-case experimental studies. Investigations at this level established the necessary and sufficient conditions for the efficacy of fear reduction methods such as desensitization and in vivo exposure, and they continue to be a major vehicle for examining the theoretical mechanisms underlying successful treatment effects (e.g., Bandura, 1977). Controlled clinical research in which these anxiety reduction methods were applied to phobic patients was then conducted, followed by long-term evaluations of the durability of treatment effects and exploration of variations in their practical implementation and cost-effectiveness (e.g., group versus individual therapy, the use of nurse therapists, and even self-administered treatment; Marks, 1981).

The influence between different levels in this framework of the progression of therapy research is not unidirectional: The path between the laboratory and the clinic is a two-way one. Thus in the case of the development of fear reduction treatments, Wolpe's early application of desensitization to patients in an uncontrolled clinical series encouraged experimental research that later resulted in the refinement of more effective methods of in vivo exposure. Another example of this process can be seen in the development of cognitive-behavioral treatment for depression (Beck et al., 1979). Beck, a psychiatrist, developed this approach in his clinical practice with depressed patients. Subsequent analogue and then clinical outcome research studies in this and other countries confirmed the efficacy of his methods (Rachman and Wilson, 1980). The success of Beck's treatment led to a decision by the Natonal Institute of Mental Health to fund a major, multicenter comparative outcome study in which cognitive-behavioral treatment is compared with interpersonal psychotherapy and pharmacotherapy in the treatment of depression (Waskow et al., 1979). Concomitant with these developments and with another much-publicized practitioner's methods--Ellis's (1962) rational-emotive therapy--Beck's clinical findings have spurred basic experimental research on cognitive processes in anxiety and depressive disorders. The recent outpouring of research on attributional mechanisms in depression, on the effect of mood on memory, and on specific cognitive correlates of phobic reactions owes much to the impact of advances in clinical treatment (e.g., Craighead et al., 1979; Kihlstrom and Nasby, 1981; Metalsky and Abramson, 1981).

It is clear that the biggest deficits in the framework of therapy research presented in Figure 1 are at the level of clinical research with patient populations, long-term follow-ups, and adequate field testing of the disseminability and implementation of research-based treatment methods in general clinical practice. Although a recent analysis of the status of clinical research in behavior therapy indicates that it is increasing (Agras and Berkowitz, 1980), controlled clinical studies still lag behind basic laboratory research.

Reasons for the relative lack of clinical research are discussed by Agras et al. (1979) and Wilson (1982). To summarize, much of the research done in the United States is based in graduate departments of psychology in universities. These settings have not been well suited to clinical research. Aside from the conceptual biases of many academic psychologists against applied research, there are serious practical obstacles that must be overcome. Controlled clinical research is expensive and time-consuming. Appropriate funding over sufficiently lengthy periods of time is often a problem. The necessary clinical and administrative support for continuing contact with patients throughout the year is frequently unavailable. Serious clinical disorders do not recognize the semester breaks or summer holidays that are so much a part of departmental functioning in universities. The competitive pressures within academia today militate against long-term, risky research. Graduate students need to publish to compete for disappearing academic and research appointments. Young faculty must bolster the curriculum vitae to obtain tenure. The contingencies, at least in part, support completion of short-term laboratory studies. Finally, clinical research is complex and multifaceted; it necessitates collaboration and a team effort with time-consuming demands of proper assessment and the conduct of therapy itself. The research team will usually benefit by being interdisciplinary in makeup, another practical problem for some psychological researchers.

In the future it will be necessary to attend to the practical and administrative problems that interfere with our ability to complete the controlled clinical investigations that provide the bridge between more basic research and clinical practice. It is research at this level that will help to close the continuing gap between the laboratory and the clinic.

REFERENCES

Agras, W. S., and R. Berkowitz
1980 "Clinical research in behavior therapy: half way there?" Behavior Therapy 11:472-487.

Agras, W. S., A. E. Kazdin, and G. T. Wilson
1979 Behavior Therapy: Towards an Applied Clinical Science. San Francisco: Freeman.

Ayllon, T., and N. H. Azrin
1965 "The measurement and reinforcement of behavior of psychotics." Journal of the Experimental Analysis of Behavior 8:357-383.

Bandura, A.
1969 Principles of Behavior Modification. New York: Holt, Rinehart & Winston.
1977 Social Learning Theory. Englewood-Cliffs, N.J.: Prentice Hall.

Barlow. D. H., and B. Wolfe
1980 "Behavioral approaches to anxiety disorders: a report on the NIMH-SUNY research conference." Journal of Consulting and Clinical Psychology 49:448-454.

Beck, A. T., A. J. Rush, B. F. Shaw, and G. Emery
1979 Cognitive Therapy of Depression. New York: Guilford Press.

Craighead, E., W. H. Kimball, and P. J. Rehak
1979 "Mood changes, physiological responses, and self-statements during social rejection imagery." Journal of Consulting and Clinical Psychology 47:385-396.

Dember, W.
1974 "Motivation and the cognitive revolution." American Psychologist 29:161-168.

Ellis, A.
1962 Reason and Emotion in Psychotherapy. New York: Lyle Stuart.

Emmelkamp, P. M. G., and A. C. M. Kuipers
1979 "Agoraphobia: a follow-up study four years after treatment." British Journal of Psychiatry 134:352-355.

Eysenck, H. J.
1959 "Learning theory and behaviour therapy." British Journal of Mental Science 105:61-75.

Fairweather, G. W., D. H. Sanders, D. L. Cressler, and H. Maynard
1969 Community Life for the Mentally Ill. Chicago: Aldine.

Fisher, S., and R. P. Greenberg
1977 The Scientific Credibility of Freud's Theories and Therapy. New York: Basic Books.

Gelder, M. G., J. H. J. Bancroft, D. Gath, D. W. Johnston, A. M. Mathews, and P. M. Shaw
1973 "Specific and non-specific factors in behaviour therapy." British Journal of Psychiatry 123:445-462.

Hersen, M., and D. H. Barlow
1976 Single-Case Experimental Designs: Strategies for Studying Behavior Change. New York: Pergamon Press.

Jacob, R., H. Kraemer, and W. S. Agras
1977 "Relaxation therapy in the treatment of hypertension." Archives of General Psychiatry 34:1417-1427.

Jacobson, E.
1938 Progressive Relaxation. Chicago: University of Chicago Press.

Kanfer, F. H., and J. S. Phillips
1970 Learning Foundations of Behavior Therapy. New York: Wiley.

Kazdin, A. E.
1978a History of Behavior Modification. Baltimore: University Park Press.
1978b "The application of operant techniques in treatment, rehabilitation, and education." In S. L. Garfield and A. E. Bergin, eds., Handbook of Psychotherapy and Behavior Change. 2nd ed. New York: Wiley.

Kazdin, A. E., and G. A. Smith
1979 "Covert conditioning: a review and evaluation." Advances in Behaviour Research and Therapy 2:57-98.

Kazdin, A. E., and G. T. Wilson
1978 Evaluation of Behavior Therapy: Issues, Evidence and Research Strategies. Cambridge, Mass.: Ballinger.

Kendall, P., and S. Hollon, eds.
1979 Cognitive-Behavioral Interventions: Theory, Research, and Procedures. New York: Guilford Press.

Kihlstrom, J., and W. Nasby
1981 "Cognitive tasks in clinical assessment: an exercise in applied psychology." In P. Kendall and S. Hollon, eds., Assessment Strategies for

Cognitive-Behavioral Interventions. New York: Academic Press.

Kimble, G.
1961 Conditioning and Learning. New York: Appleton-Century-Crofts.

Lang, P. J.
1969 "The mechanics of desensitization and the laboratory study of fear." In C. M. Franks, ed., Behavior Therapy: Appraisal and Status. New York: McGraw-Hill.
1979 "A bio-informational theory of emotional imagery." Psychophysiology 16:495-512.

Leitenberg, H.
1976 "Behavioral approaches to treatment of neuroses." In H. Leitenberg, ed., Handbook of Behavior Modification and Behavior Therapy. Englewood Cliffs, N.J.: Prentice-Hall.

Levis, D., and N. Hare
1977 "A review of the theoretical rationale and empirical support for the extinction approach of implosive (flooding) therapy." In M. Hersen, R. Eisler, and P. Miller, eds., Progress in Behavior Modification. Volume 4. New York: Academic Press.

Maccoby, N., J. Farquhar, P. D. Wood, and J. Alexander
1977 "Reducing the risk of cardiovascular disease: effects of a community-based campaign on knowledge and behavior." Journal of Community Health 3:100-114.

Mahoney, M. J.
1974 Cognition and Behavior Modification. Cambridge, Mass.: Ballinger.

Marks, I. M.
1971 "Phobic disorders four years after treatment: a prospective follow-up." British Journal of Psychiatry 118:683-688.
1981 Cure and Care of Neuroses. New York: Wiley.

Marks, I. M., R. S. Hallam, J. Connolly, and R. Philpott
1977 Nursing in Behavioural Therapy. London: The Royal College of Nursing of the United Kingdom.

Marks, I., R. Stern, D. Mawson, J. Cobb, and R. McDonald
1980 "Clomipramine and exposure for obsessive-compulsive rituals: I." British Journal of Psychiatry 136:1-25.

McPherson, F. M , L. Brougham, and S. McLaren
1980 "Maintenance of improvement in agoraphobic patients treated by behavioural methods--a

four-year follow-up." Behaviour Research and Therapy 18:150-152.

Meichenbaum, D.
1977 Cognitive Behavior Modification. New York: Plenum.

Meichenbaum, D., and J. Asarnow
1979 "Cognitive-behavior modification and meta cognitive development: implications for the classroom." In P. E. Kendall and S. D. Hollon, eds., Cognitive-Behavioral Interventions: Theory, Research and Procedures. New York: Academic Press.

Melamed, B., and L. Siegel
1980 Behavioral Medicine. New York: Springer.

Metalsky, G. I., and L. Y. Abramson
1981 "Attributional styles: toward a framework for conceptualization and assessment." In P. Kendall and S. Hollon, eds., Assessment Strategies for Cognitive-Behavioral Interventions. New York: Academic Press.

Miller, N. E.
1969 "Learning of visceral and glandular responses." Science 163:434-445.

Mills, H. L., W. S. Agras, D. H. Barlow, and J. R. Mills
1973 "Compulsive rituals treated by response prevention." Archives of General Psychiatry 28:524-529.

Mowrer, O., and W. Mowrer
1938 "Enuresis: a method for its study and treatment." American Journal of Orthopsychiatry 8:436-459.

Munby, M., and D. W. Johnston
1980 "Agoraphobia: the long-term follow-up of behavioural treatment." British Journal of Psychiatry 137:418-427.

Paul, G. L.
1969 "Behavior modification research: design and tactics." In C. M. Franks, ed., Behavior Therapy: Appraisal and Status. New York: McGraw-Hill.

Paul, G. L., and R. J. Lentz
1977 Psychological Treatment of Chronic Mental Patients. Cambridge, Mass.: Harvard University Press.

Rachman, S., and R. Hodgson
1980 Obsessions and Compulsions. Englewood Cliffs, N.J.: Prentice-Hall.

Rachman, S., and G. T. Wilson
1980 The Effects of Psychological Therapy. Oxford: Pergamon Press.

Rosenthal, T. L.
1982 "Social learning theory and behavior therapy." In G. T. Wilson and C. M. Franks, eds., Contemporary Behavior Therapy: Conceptual and Empirical Foundations. New York: Guilford Press.

Rosenthal, T., and A. Bandura
1978 "Psychological modeling: theory and practice." In S. L. Garfield and A. E. Bergin, eds., Handbook of Psychotherapy and Behavior Change. New York: Wiley.

Ross, A.
1981 Child Behavior Therapy. New York: Wiley.

Ross, J.
1980 "The use of former phobics in the treatment of phobias." American Journal of Psychiatry 137:715-717.

Silver, B. V., and E. B. Blanchard
1978 "Biofeedback and relaxation training in the treatment of psychophysiological disorders: or are the machines really necessary?" Journal of Behavioral Medicine 1:217-239.

Skinner, B. F.
1953 Science and Human Behavior. New York: Macmillan.

Solomon, R., L. Kamin, and L. Wynne
1953 "Traumatic avoidance learning: the outcome of several extinction procedures with dogs." Journal of Abnormal and Social Psychology 48:291-302.

Stuart, R. B., ed.
1977 Behavioral Self-Management. New York: Brunner/Mazel.

Turkewitz, H., K. D. O'Leary, and M. Ironsmith
1975 "Generalization and maintenance of appropriate behavior through self-control." Journal of Consulting and Clinical Psychology 43:577-583.

Waskow, I., S. Hadley, M. Parloff, and J. Autrey
1979 Psychotherapy of Depression Collaborative Research Program. Unpublished manuscript. Clinical Research Branch, National Institute of Mental Health, Rockville, Md.

Wilson, G. T.
1982 "Behavior therapy for adults: application and outcome." In G. T. Wilson and C. M. Franks, eds., Contemporary Behavior Therapy: Conceptual and Empirical Foundations. New York: Guilford Press.

Wilson, G. T., and K. Brownell
1980 "Behavior therapy for obesity: an evaluation of treatment outcome." Advances in Behaviour Research and Therapy 3:49-86.

Wilson, G. T., and I. M. Evans
1977 "The therapist-client relationship in behavior therapy." In R. S. Gurman and A. M. Razin, eds., The Therapist's Contribution to Effective Psychotherapy: An Empirical Approach. New York: Pergamon Press.

Wilson, G. T., and K. D. O'Leary
1980 Principles of Behavior Therapy. Englewood Cliffs, N.J.: Prentice-Hall.

Wolpe, J.
1958 Psychotherapy by Reciprocal Inhibition. Stanford, Calif.: Stanford University Press.

CONTRIBUTORS

L. D. BRAIDA, Center for Communications Sciences, Research Laboratory of Electronics, Massachusetts Institute of Technology
PATRICIA A. CARPENTER, Department of Psychology, Carnegie-Mellon University
RICHARD CONTRADA, Graduate Center, City University of New York
PHILIP E. CONVERSE, Institute for Social Research, University of Michigan
TOM N. CORNSWEET, School of Social Sciences, University of California, Irvine
ROY G. D'ANDRADE, Department of Anthropology, University of California, San Diego
N. I. DURLACH, Center for Communications Sciences, Research Laboratory of Electronics, Massachusetts Institute of Technology
HEINZ EULAU, Department of Political Science, Stanford University
DAVID L. FEATHERMAN, Department of Sociology, University of Wisconsin
DAVID C. GLASS, Graduate Center, City University of New York
DAVID M. GREEN, Laboratory of Psychophysics, Department of Psychology and Social Relations, Harvard University
JAMES J. HECKMAN, Economics Research Center, National Opinion Research Center, and Department of Economics, University of Chicago
MARCEL A. JUST, Department of Psychology, Carnegie-Mellon University
DAVID S. KRANTZ, Department of Medical Psychology, Uniformed Services University of the Health Sciences
HERSCHEL LEIBOWITZ, Department of Psychology, Pennsylvania State University

ALVIN LIBERMAN, Haskins Laboratories, New Haven, Connecticut

R. DUNCAN LUCE, Department of Psychology and Social Relations, Harvard University

JANE MENKEN, Office of Population Research, Princeton University

ROBERT T. MICHAEL, Economics Research Center, National Opinion Research Center, and Department of Education, University of Chicago

NEAL E. MILLER, Rockefeller University

WARREN E. MILLER, Department of Political Science, Arizona State University, and Department of Political Science, University of Michigan

KATHERINE NELSON, Graduate Center, City University of New York

ROBERT McC. NETTING, Department of Anthropology, University of Arizona

RICHARD PEW, Bolt Beranek & Newman, Cambridge, Massachusetts

CARL SHERRICK, Department of Psychology, Princeton University

JUDITH M. TANUR, Department of Sociology, State University of New York, Stony Brook

JAMES TRUSSELL, Office of Population Research, Princeton University

G. TERENCE WILSON, Graduate School of Applied and Professional Psychology, Rutgers University